



---

## Neural Signatures of Trust during Human-Automation Interactions

Frank Krueger  
GEORGE MASON UNIVERSITY

---

04/01/2016  
Final Report

DISTRIBUTION A: Distribution approved for public release.

Air Force Research Laboratory  
AF Office Of Scientific Research (AFOSR)/ RTA2  
Arlington, Virginia 22203  
Air Force Materiel Command

<b>REPORT DOCUMENTATION PAGE</b>					<i>Form Approved</i> OMB No. 0704-0188	
<small>The public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing the burden, to the Department of Defense, Executive Service Directorate (0704-0188). Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to any penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.</small>						
<b>PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ORGANIZATION.</b>						
<b>1. REPORT DATE (DD-MM-YYYY)</b> 30-03-2016		<b>2. REPORT TYPE</b> Final			<b>3. DATES COVERED (From - To)</b> 01-01-2013 to 12-31-2015	
<b>4. TITLE AND SUBTITLE</b> Neural Signatures of Trust during Human-Automation Interactions				<b>5a. CONTRACT NUMBER</b> FA9550-13-1-0017		
				<b>5b. GRANT NUMBER</b>		
				<b>5c. PROGRAM ELEMENT NUMBER</b>		
<b>6. AUTHOR(S)</b> Krueger, Frank				<b>5d. PROJECT NUMBER</b>		
				<b>5e. TASK NUMBER</b>		
				<b>5f. WORK UNIT NUMBER</b>		
<b>7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)</b> George Mason University 4400 University Drive Fairfax, VA 22030-4422					<b>8. PERFORMING ORGANIZATION REPORT NUMBER</b>	
<b>9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)</b> Air Force Office of Scientific Research Program Officer - Trust and Influence 875 N. Randolph St. Arlington, VA 22203					<b>10. SPONSOR/MONITOR'S ACRONYM(S)</b> AFOSR	
					<b>11. SPONSOR/MONITOR'S REPORT NUMBER(S)</b>	
<b>12. DISTRIBUTION/AVAILABILITY STATEMENT</b> Approved for Public Release						
<b>13. SUPPLEMENTARY NOTES</b> N/A						
<b>14. ABSTRACT</b> The objective of this proposal was to investigate the similarities and differences of the neural systems of human-automation trust (HAT) and human-human trust (HHT) in a series of three studies that combined an X-ray luggage-screening task with functional magnetic resonance imaging by manipulating the reliability of advice from a human or automated luggage inspector framed as experts. HAT and HHT were measured as the acceptance rates of advice either giving by the machine or the human agent. Comparing HAT with HHT, those studies provide first neural evidence that reliable (study 1) and unreliable (false alarm [study2] and misses [study 3]) human-automation interactions evoke unique brain activation patterns linked with the reward network for reinforcement learning (e.g., dorsal striatum head, ventromedial prefrontal cortex), the mentalizing network for evaluating personal characteristics and traits (e.g., precuneus, temporoparietal junction), and the salience network for interoception (e.g., insula, anterior cingulate cortex). The findings are relevant to the Air Force Office of Scientific Research's mission aimed at investing in the discovery of the foundational concepts of trust building and trust calibration during complex human-machine interactions.						
<b>15. SUBJECT TERMS</b> trust, human-human trust, human-automation trust, brain, functional magnetic resonance imaging						
<b>16. SECURITY CLASSIFICATION OF:</b> a. REPORT b. ABSTRACT c. THIS PAGE			<b>17. LIMITATION OF ABSTRACT</b>		<b>18. NUMBER OF PAGES</b>	
					<b>19a. NAME OF RESPONSIBLE PERSON</b>	
					<b>19b. TELEPHONE NUMBER (Include area code)</b>	

Reset

## INSTRUCTIONS FOR COMPLETING SF 298

**1. REPORT DATE.** Full publication date, including day, month, if available. Must cite at least the year and be Year 2000 compliant, e.g. 30-06-1998; xx-06-1998; xx-xx-1998.

**2. REPORT TYPE.** State the type of report, such as final, technical, interim, memorandum, master's thesis, progress, quarterly, research, special, group study, etc.

**3. DATES COVERED.** Indicate the time during which the work was performed and the report was written, e.g., Jun 1997 - Jun 1998; 1-10 Jun 1996; May - Nov 1998; Nov 1998.

**4. TITLE.** Enter title and subtitle with volume number and part number, if applicable. On classified documents, enter the title classification in parentheses.

**5a. CONTRACT NUMBER.** Enter all contract numbers as they appear in the report, e.g. F33615-86-C-5169.

**5b. GRANT NUMBER.** Enter all grant numbers as they appear in the report, e.g. AFOSR-82-1234.

**5c. PROGRAM ELEMENT NUMBER.** Enter all program element numbers as they appear in the report, e.g. 61101A.

**5d. PROJECT NUMBER.** Enter all project numbers as they appear in the report, e.g. 1F665702D1257; ILIR.

**5e. TASK NUMBER.** Enter all task numbers as they appear in the report, e.g. 05; RF0330201; T4112.

**5f. WORK UNIT NUMBER.** Enter all work unit numbers as they appear in the report, e.g. 001; AFAPL30480105.

**6. AUTHOR(S).** Enter name(s) of person(s) responsible for writing the report, performing the research, or credited with the content of the report. The form of entry is the last name, first name, middle initial, and additional qualifiers separated by commas, e.g. Smith, Richard, J, Jr.

**7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES).** Self-explanatory.

**8. PERFORMING ORGANIZATION REPORT NUMBER.** Enter all unique alphanumeric report numbers assigned by the performing organization, e.g. BRL-1234; AFWL-TR-85-4017-Vol-21-PT-2.

**9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES).** Enter the name and address of the organization(s) financially responsible for and monitoring the work.

**10. SPONSOR/MONITOR'S ACRONYM(S).** Enter, if available, e.g. BRL, ARDEC, NADC.

**11. SPONSOR/MONITOR'S REPORT NUMBER(S).** Enter report number as assigned by the sponsoring/monitoring agency, if available, e.g. BRL-TR-829; -215.

**12. DISTRIBUTION/AVAILABILITY STATEMENT.** Use agency-mandated availability statements to indicate the public availability or distribution limitations of the report. If additional limitations/ restrictions or special markings are indicated, follow agency authorization procedures, e.g. RD/FRD, PROPIN, ITAR, etc. Include copyright information.

**13. SUPPLEMENTARY NOTES.** Enter information not included elsewhere such as: prepared in cooperation with; translation of; report supersedes; old edition number, etc.

**14. ABSTRACT.** A brief (approximately 200 words) factual summary of the most significant information.

**15. SUBJECT TERMS.** Key words or phrases identifying major concepts in the report.

**16. SECURITY CLASSIFICATION.** Enter security classification in accordance with security classification regulations, e.g. U, C, S, etc. If this form contains classified information, stamp classification level on the top and bottom of this page.

**17. LIMITATION OF ABSTRACT.** This block must be completed to assign a distribution limitation to the abstract. Enter UU (Unclassified Unlimited) or SAR (Same as Report). An entry in this block is necessary if the abstract is to be limited.

**Subject:**

Final Performance Report to Dr. Benjamin A. Knott

**To:**

technicalreports@afosr.af.mil, benjamin.knott@wpafb.af.mil

**Contract/Grant Title:**

Neural Signatures of Trust during Human-Automation Interactions

**Principal Investigator's (PI) name:**

Dr. Frank Kruger

**Institution's Name and Address:**

George Mason University  
Krasnow Institute for Advanced Study  
4400 University Drive  
Fairfax, VA 22030-4422

**Contract/Grant #:**

FA9550-13-1-0017

**Reporting Period:**

01/01 2013 to 12/31 2015

## 1. Executive Summary

Well-calibrated human-automation trust (HAT) is an essential ingredient for efficiency, communication, and safety in complex human-automation interactions. A dichotomy between HAT and human-human trust (HHT) has been proposed: some scholars argue that HAT and HHT are fundamentally different due to initial perception and lack of intention on the part of automation, while others claim that HAT and HHT are equal, since similar social interactions as between humans can be elicited when automation is designed to be human-like. Although, recent behavioral research has provided evidence for both accounts and a plethora of neural evidence for HHT already exists; however, the underlying neural signatures for HAT and its relationship to HHT are still unexplored. Behavioral measures alone are unlikely to allow one to distinguish between HHT and HAT, because the same behavioral outcome can be associated with very different underlying neural mechanisms. Assessing both performance and brain function can provide more information than either alone. The objective of this proposal was to investigate the similarities and differences of the neural systems of HAT and HHT in a series of three studies that combined a behavioral X-ray luggage-screening task with functional magnetic resonance imaging (fMRI) and manipulated reliabilities of advice (unknown to the participants) as the key feature for HAT and HHT interactions. Healthy participants were asked to search for knives hidden in densely cluttered X-ray images of luggage after receiving advice (presence or absence of a knife) from a human or automated luggage inspector (framed as experts). HAT and HHT were measured as the acceptance rates of advice either giving by the machine or human agent. By adopting a comprehensive, interdisciplinary research program including scientists from social cognitive neuroscience, psychology, and human factors, we accomplished the overall objective of this proposal by pursuing the following three specific aims:

**Aim #1: Neural signatures of HAT based on reliable human-automation interactions.** In study 1, participants performed the security screening task and decided whether to search or clear the luggage after receiving advice from a human or automated luggage inspector with a manipulated reliability of 90%. HHT was initially lower than HAT, probably due to the preconceived notions of automation being perfect. However, over time differences between HHT and HAT disappeared based on a higher degree of confidence toward the human adviser to perform the task based on the received feedback. This reinforcement learning process was mirrored by activations in reward-sensitive brain regions, including the dorsal striatum and ventromedial prefrontal cortex. In summary, comparing HHT and HAT study 1 provided the

first neural evidence showing how automation bias mediates these types of trust, thus leading to behavioral differences in the context of advice taking.

**Aim #2: Neural signatures of HAT based on unreliable human-automation interactions due to high false alarm rates.** In study 2, participants completed the X-ray luggage-screening task by either rejecting or accepting bad or good advice from either a machine or human inspector with a manipulated reliability of 60% (false alarm rate). Unreliable advice decreased performance overall. HHT was lower than HAT during bad advice, presumably due to reevaluation of expectations arising from association of dispositional credibility for each agent. Trust differences engaged brain regions associated with the mentalizing network for evaluating personal characteristics and traits (precuneus, posterior cingulate cortex, temporoparietal junction) and the salience network for interoception (posterior insula). Posterior insula and left precuneus were the drivers of the HHT network that were reciprocally connected to each other and also projected to all other regions. In summary, study 2 revealed insights into the neural underpinnings of HAT and HHT associated with unreliable advice utilization due to high false alarm rates.

**Aim #3: Neural Signatures of HAT based on unreliable human-automation interactions due to high miss rates (60%).** In study 3, participants performed the X-ray luggage-screening task by either accepting or rejecting good or bad advice from either a human or a machine inspector with a manipulated reliability 60% (miss rate) of. HAT decreased more than HHT over time, possibly due to high expectations of reliable advice from a machine and changes in attention allocation due to miss errors. Brain areas involved with the salience and mentalizing networks, as well as sensory processing involved with attention were less active for HAT as for HHT. The HAT network consisted of attentional modulation of sensory information with the lingual gyrus as the driver during the decision phase and the fusiform gyrus as the driver during the feedback phase of the task. In summary, study 3 expanded on the existing literature by showing how misses degrade HAT in comparison to HHT, which is represented in brain regions involved in salience detection and self-processing with perceptual integration.

The performed studies are innovative, because they were among the first directly to examine and compare the neural signatures of HAT (and its relationship to HHT) in the context of human-automation performance applying a multi-disciplinary approach. The findings have

significant implications for society because of progressions in technology and increased interactions with machines. Moreover, those findings are relevant to the Air Force Office of Scientific Research's mission aimed at fostering innovative research and enhancing the Air Force's impact on policies and operations related to national security by investing in the discovery of the foundational concepts of trust building and trust calibration during complex human-machine interactions. Overall, the successful completion of this project resulted in two substantive project outcomes: first, a significant increase in our knowledge about the underlying neural circuits of HAT calibration during complex human-automation interactions and second, the laboratory results provide a methodology and rationale for exploring HAT in field research and for developing transformative novel theories and models.

## **2. Personnel Supported:**

PI: Dr. Frank Krueger

Co-PI: Dr. Raja Parasuraman passed away during the last year of the project.

Graduate student: Kimberly Goodyear

## **3. Publications:**

Findings of study 1 were submitted as an abstract to the 21st Annual Meeting of the Cognitive Neuroscience Society (Boston, MA; April 5-8, 2014):

Title: How automation bias influences human-human and human-automation trust: An fMRI study

Authors: Goodyear K, Bowman A, Chernyak S, De Visser E, Parasuraman R, Krueger F.

Findings of study 2 were submitted as an abstract to the Society for Social Neuroscience Annual Meeting (Chicago, IL; October 16, 2015):

Title: Comparisons of advice utilization during human and machine agent interactions: a functional magnetic resonance imaging and effective connectivity study

Authors: Goodyear K, Parasuraman R, Chernyak S, Madhavan P, Deshpande G, Krueger F.

The research effort for this project culminated in the production of one dissertation. In April 2006, Kimberly S. Goodyear will defend her dissertation entitled "The neural basis of advice utilization During human and machine agent interactions" to the graduate faculty of George Mason University in partial fulfillment of the requirements for the degree of Doctor of Philosophy Neuroscience. The dissertation includes the findings from study 1 and study 2 (see attachment). The PI of the research project will act as the Dissertation Director.

Moreover, a manuscript entitled “Advice utilization during human and machine interactions: an fMRI and effective connectivity study” based on the findings of study 2 is currently under review as an original research article in the journal “Frontiers in Human Neuroscience”:

Authors: Kimberly Goodyear, Raja Parasuraman, Sergey Chernyak, Poornima Madhavan, Gopikrishna Deshpande, Frank Krueger

Author Contributions: K.G. and S.C. acquired the data for analysis. K.G., R.P. and F.K. contributed to the conception of the design. K.G., R.P., S.C., P.M., G.D. and F.K. contributed to interpretation of the data. K.G., R.P., S.C., P.M., G.D. and F.K. contributed to drafting of the work and revising it critically. K.G., R.P., S.C., P.M., G.D. and F.K. approved the final version to be published. K.G., R.P., S.C., P.M., G.D. and F.K. agreed to be accountable for all aspects of the work.

Abstract: With new technological advances, advice can come from different sources such as machines or humans, but how individuals respond to such advice and the neural correlates involved need to be better understood. We combined functional MRI and multivariate Granger causality analysis with an X-ray luggage-screening task to investigate the neural basis and corresponding effective connectivity involved with advice utilization from agents framed as experts. Participants were asked to accept or reject good or bad advice from a human or machine agent with manipulated reliability (high false alarm rate). We showed that unreliable advice decreased performance overall and participants interacting with the human agent had a greater depreciation of advice utilization during bad advice. These differences in advice utilization can be due to reevaluation of expectations arising from association of dispositional credibility for each agent. We demonstrated that differences in advice utilization engaged brain regions associated with evaluation of personal characteristics and traits (precuneus, posterior cingulate cortex, temporoparietal junction) and interoception (posterior insula). We found that the right posterior insula and left precuneus were the drivers of the advice utilization network that were reciprocally connected to each other and also projected to all other regions. Our behavioral and neuroimaging results have significant implications for society because of progressions in technology and increased interactions with machines.

Finally, another manuscript entitled “An fMRI and effective connectivity study investigating miss errors during advice utilization from human and machine agents” based on the findings of study 3 is currently under review as an original research article in the journal “Social Neuroscience”:

Authors: Kimberly Goodyear, Raja Parasuraman, Sergey Chernyak, Ewart de Visser, Poornima Madhavan, Gopikrishna Deshpande, Frank Krueger

Author Contributions: K.G. and S.C. acquired the data for analysis. K.G., R.P. and F.K. contributed to the conception of the design. K.G., R.P., S.C., P.M., G.D. and F.K. contributed to interpretation of the data. K.G., R.P., S.C., E.D.V., P.M., G.D. and F.K. contributed to drafting of the work and revising it critically. K.G., R.P., S.C., E.D.V., P.M., G.D. and F.K. approved the final version to be published. K.G., R.P., S.C., E.D.V., P.M., G.D. and F.K. agreed to be accountable for all aspects of the work.

Abstract. As society becomes more reliant on machines and automation, understanding how people utilize advice is a necessary endeavor. Our objective was to reveal the underlying neural mechanisms during advice utilization from expert human and machine agents with fMRI and multivariate Granger causality analysis. During an X-ray luggage-screening task, participants accepted or rejected good or bad advice from either the human or machine agent framed as experts with manipulated reliability (high miss rate). We showed that the machine-agent group decreased their advice utilization compared to the human-agent group and these differences in behaviors during advice utilization could be accounted for by high expectations of reliable advice and changes in attention allocation due to miss errors. Brain areas involved with the salience and mentalizing networks, as well as sensory processing involved with attention, were recruited during the task and the advice utilization network consisted of attentional modulation of sensory information with the lingual gyrus as the driver during the decision phase and the fusiform gyrus as the driver during the feedback phase. Our findings expand on the existing literature by showing that misses degrade advice utilization, which is represented in a neural network involving salience detection and self-processing with perceptual integration.

#### **4. Change in AFOSR program manager, if any:**

Dr. Benjamin Knott replaced Dr. Joseph Lyons on August 1st, 2013 as the Program Officer for the Trust and Influence portfolio.

#### **5. Extensions granted or milestones slipped, if any:**

None

#### **6. Include any new discoveries, inventions, or patent:**

None

#### **7. Disclosures during this reporting period (if none, report none):**

None

THE NEURAL BASIS OF ADVICE UTILIZATION DURING HUMAN AND  
MACHINE AGENT INTERACTIONS

by

Kimberly S. Goodyear  
A Dissertation  
Submitted to the  
Graduate Faculty  
of  
George Mason University  
in Partial Fulfillment of  
The Requirements for the Degree  
of  
Doctor of Philosophy  
Neuroscience

Committee:

_____	Dr. Frank Krueger, Dissertation Director
_____	Dr. Gopikrishna Deshpande, Committee Member
_____	Dr. Kevin McCabe, Committee Member
_____	Dr. William Kennedy, Committee Member
_____	Dr. David Wu, Interim Director, Krasnow Institute for Advanced Study
_____	Dr. Donna M. Fox, Associate Dean, Office of Student Affairs & Special Programs, College of Science
_____	Dr. Peggy Agouris, Dean, College of Science

Date: \_\_\_\_\_ Spring Semester 2016  
George Mason University  
Fairfax, VA

The Neural Basis of Advice Utilization During Human and Machine Agent Interactions  
A Dissertation submitted in partial fulfillment of the requirements for the degree of  
Doctor of Philosophy at George Mason University

by

Kimberly S. Goodyear  
Bachelor of Science  
San Diego State University, 2005

Director: Frank Krueger, Professor  
Department of Molecular Neuroscience

Spring Semester 2016  
George Mason University  
Fairfax, VA

This work is liscened under a Copyright 2016 by Kimberly Goodyear. All Rights Reserved.

## **DEDICATION**

I dedicate this dissertation to my loving grandfather, Clyde James Goodyear (September 10, 1914 – June 12, 2007). I kept my promise to you to “make something of myself.”

## **ACKNOWLEDGEMENTS**

I would like to thank my family for always being there for me. My parents, Clyde and Soon provided me with unconditional love and support. Their weekly talks kept my spirits high throughout my entire time as a graduate student. My brother, C.J., who always stood by side through all the challenges I have faced. I give them so much gratitude for their encouragement and reassurance.

I cannot express into words how thankful I am for the support I received from David Chavanne and our dog, Sadie. Without their love and endless assurance, this dissertation would have been almost unattainable. David's validation and confidence in my work kept my mind focused and my resolve intact. Thank you for always standing by my side.

Thank you to my advisor, Frank Krueger and committee members, Kevin McCabe, William Kennedy, Gopikrishna Deshpande for the help and support of this dissertation. I would also like to thank a past committee member, Raja Parasuraman for his contribution to this dissertation. His involvement in my research was truly invaluable and he will be missed greatly.

## TABLE OF CONTENTS

	Page
List of Tables .....	vii
List of Figures .....	viii
Abstract .....	ix
Chapter One: General Introduction.....	1
1.1 Advice Utilization .....	1
1.2 Errors .....	3
1.3 Overview of the Studies .....	7
Chapter Two: The Impact of False Alarms on Advice Utilization .....	9
2.1 Abstract .....	9
2.2 Introduction .....	11
2.3 Methods .....	14
2.4 Results .....	23
2.5 Discussion .....	34
2.6 References .....	41
Chapter Three: The Impact of Misses on Advice Utilization .....	47
3.1 Abstract .....	47
3.2 Introduction .....	49
3.3 Methods .....	52
3.4 Results .....	58
3.5 Discussion .....	70
3.6 References .....	76
Chapter Four: General Discussion .....	80
4.1 Behavioral Results.....	80
4.2 fMRI Results .....	85
4.3 Future Directions and Conclusions .....	86
Appendix A: False Alarms.....	88

A.1 Experimental Setup .....	88
A.2 Human and Machine Agent Descriptions .....	90
A.3 Brain Regions Associated with the Main Effect of Advice .....	91
A.4 Schematic Illustrating the Effective Connectivity Analysis Pipeline .....	92
A.5 Behavioral Results for Decision Phase .....	93
A.6 Results for the Confidence Ratings .....	94
A.7 Descriptive Statistics for Psychological Control Measures .....	95
Appendix B: Misses .....	96
B.1 Experimental Setup.....	96
B.2 Effective Connectivity Analysis .....	97
B.3 Schematic Illustrating the Effective Connectivity Analysis Pipeline.....	98
B.4 Descriptive Statistics for Psychological Control Measures.....	100
B.5 Results for the Decision Phase .....	101
B.6 Confidence Ratings Results.....	102
B.7 Appendix B References .....	103
References.....	104

## LIST OF TABLES

Table	Page
Table 1. False Alarm Brain Regions.....	32
Table 2. False Alarm Granger Causality Analysis.....	34
Table 3. Miss Brain Regions.....	65
Table 4. Miss Granger Causality Analysis .....	70

## LIST OF FIGURES

Figure	Page
Figure 1. False Alarm Behavioral Results .....	25
Figure 2. False Alarm Rating Results .....	27
Figure 3. False Alarm Brain Activations for Decision Phase .....	29
Figure 4. False Alarm Activation Patterns During Decision Phase .....	30
Figure 5. False Alarm Brain Activations During Feedback Phase .....	31
Figure 6. False Alarm Results for Multivariate Granger Causality Analysis .....	33
Figure 7. Miss Behavioral Results .....	59
Figure 8. Miss Rating Results .....	62
Figure 9. Miss Brain Activations During Decision Phase .....	64
Figure 10. Miss Activation Patterns During Decision Phase .....	65
Figure 11. Miss Brain Activations During Feedback Phase .....	67
Figure 12. Miss Brain Activation Patterns During Feedback Phase .....	68
Figure 13. Miss Results for Multivariate Granger Causality Analysis .....	69

## **ABSTRACT**

### **THE NEURAL BASIS OF ADVICE UTILIZATION DURING HUMAN AND MACHINE AGENT INTERACTIONS**

Kimberly S. Goodyear, Ph.D.

George Mason University, 2016

Dissertation Director: Dr. Frank Krueger

Understanding how individuals utilize advice from humans and machines has become progressively more pertinent as technological advances have pervaded our society. With an increasing shift towards relying on automation, the necessity to understand the complex interactions that exist between humans and automation has emerged. This thesis examines the behavioral, cognitive and neural mechanisms involved with advice utilization from human and machine agents framed as experts. A series of two studies were implemented that consisted of an X-ray luggage-screening task with functional magnetic resonance imaging and effective connectivity analysis. To assess advice taking differences between human and machines across both studies, the agents' reliability was manipulated with high error rates. To fully ascertain how individuals respond to unreliable advice, the focus of Chapter Two was on false alarms, while in Chapter Three the focus was on misses. In each study, we demonstrated that there were unique

behavioral responses and brain activation patterns, but in both studies participant performance levels declined overall. In Chapter Two, we showed that participants interacting with the human agent had a greater depreciation of advice utilization during bad advice and there was activation in brain regions associated with evaluation of personal characteristics, traits and interoception. In addition, the effective connectivity analysis revealed that the right posterior insula and left precuneus were the drivers of the network that were reciprocally connected to each other and also projected to all other regions (right precuneus, posterior cingulate cortex, rostrolateral prefrontal cortex and posterior temporoparietal junction). In Chapter Three, we demonstrated that advice utilization decreased more for the machine-agent group and brain areas involved with the salience and mentalizing networks, as well as sensory processing involved with attention, were recruited during the task. The effective connectivity analysis showed that the lingual gyrus was the driver during the decision phase that projected to all other target regions (anterior cingulate cortex, precuneus and cuneus) and the fusiform gyrus was the driver during the feedback phase that sent an output to the inferior parietal lobule. The contribution of this thesis is a greater comprehension of the decision-making processes involved during advice taking, which may serve as a building block for uncovering the different factors involved with human-machine interactions.

## **CHAPTER ONE: GENERAL INTRODUCTION**

The prevalence of new technology in our society today has created an increased reliance on automation and with progressions in mechanization and automated aids, this allows for a heuristic to decrease human workload for manual labor (Mosier, Skitka, Heers, & Burdick, 1998; Parasuraman & Riley, 1997). For example, in 2013, the Federal Aviation Administration published a report on the operational use of flight path management systems that showed that pilot interaction with automation may result in overreliance and over 50% of accidents reviewed were due to the pilot's reduced situational awareness (Federal Aviation Administration, 2013). With this shift of job roles from human to automation, understanding how individuals vary in response to automation has become more pertinent as potential issues may arise. To better comprehend the complex nature of human-machine interactions, the rest of this chapter will explore advice utilization and the effect of errors in greater detail.

### **1.1 Advice Utilization**

The ways in which individuals respond to advice can vary depending on different factors involved during those social interactions. For example, variables such as source credibility and type of advice can influence whether a person utilizes or discounts the advice given to them. Studies have demonstrated that expert advice is used more than

novice advice (Snizek, Schrah, & Dalal, 2004) and poor (inaccurate) advice is discounted more than good (accurate) advice (Yaniv & Kleinberger, 2000). A study investigating perceptions of decision aids revealed that measures of trust varied depending on the pedigree (novice vs. expert) of the human or automated aid (Madhavan & Wiegmann, 2007), revealing that advice acceptance between humans and machines differs depending on source credibility. Moreover, the authors postulate that advice utilization strategies for humans and automation may differ due to dispositional credibility and high expectations of reliable advice. In addition, the decision to accept or reject advice may be influenced by the reliability of the source. For instance, it has been shown that automation characteristics such as reliability, predictability and ability can affect how people respond to imperfect automation (Lee & See, 2004). Initial expectations of reliable advice can be altered when disconfirmation evidence about an agent's credibility is revealed. For example, a study demonstrated that initial confirmatory experience can increase how much a person follows bad advice, which ultimately impacts decision-making behaviors (Staudinger & Buchel, 2013). This phenomenon can be explained in terms of an expectation disconfirmation theory, where an expectation is a belief that someone or something will live up to what is anticipated and disconfirmation is a discrepancy in the evaluation of that expectation (Oliver, 1980).

Neuroimaging studies investigating advice taking, personal traits, dispositions and human-robot interactions have revealed the involvement with regions associated with the salience, mentalizing and central executive networks (Brosch, Schiller, Mojdehbakhsh, Uleman, & Phelps, 2013; Chaminade et al., 2012; Krach et al., 2008; Suen, Brown,

Morck, & Silverstone, 2014). Menon (2011) proposed a model involving three large-scale brain networks, including the central executive network (CEN), the salience network (SN) and the default-mode network (DMN); the CEN (dorsolateral PFC, dlPFC) has been postulated to be involved with higher-order cognitive functions such as decision-making, the SN (dorsal ACC, AI) has been implicated in saliency detection of internal and external events and the DMN (PCC, PreC) has been revealed to be associated with self-processing cognitions. A study investigating tracking of expertise for humans and algorithms found areas associated with the mentalizing network and salience networks (e.g., ACC, precuneus) during estimates of the agents' abilities (Boorman, O'Doherty, Adolphs, & Rangel, 2013). In addition, studies investigating observations of human and robot interactions (Suen et al., 2014) and inferences of mental states for humans and machines (Chaminade et al., 2012), as well as studies examining attribution of personal traits and characteristics (Cabanis et al., 2013) have shown recruitment of areas associated with large-scale brain networks. Given considerable overlap between the aforementioned neural networks, the overall aim of this thesis was to investigate the underlying mechanisms involved with advice taking from humans and machines to provide potential evidence about how individuals perceive and utilize advice from different agents.

## **1.2 Errors**

To provide a background understanding for the circumstances in which a person makes decisions during levels of uncertainty (i.e., unreliable advice), Signal Detection

Theory (SDT) can show how differences in advice utilization pertain to differences in error types (Tanner Jr & Swets, 1954). To measure the individual responses according to performance rates, there are signal absent (correct rejection [correct non-alert]), false alarm [incorrect alert]) and signal present (hit [correct alert], miss [incorrect non-alert]) distributions. Looking at the different error types (false alarms and misses) within a decision matrix allows for an even greater interpretation of the factors involved with advice utilization. For example, it has been shown that false alarms can hurt overall performance, operator compliance (agreeing when the aid indicates the target is present) and operator reliance (agreeing when the aid indicates the target is absent), while misses only affect operator reliance (Dixon, Wickens, & McCarley, 2007; Rice & McCarley, 2011). However, there are conflicting views pertaining to the overlap between compliance and reliance, which warrants further exploration on the topic (Dixon et al., 2007; Meyer, 2004). Moreover, it has been revealed that false alarms may cause operators to not respond to alerts at all, which has been coined as the “cry wolf effect,” (Breznitz, 2013; Wickens et al., 2009) and furthermore, misses may affect monitoring strategies during non-alarm periods causing a change in attention allocation strategies (Onnasch, Ruff, & Manzey, 2014).

A comprehensive review by McBride, Rogers, and Fisk (2014) determined that management of automation errors can be broken up into a framework of four variables: automation characteristics (e.g., reliability), person factors (e.g., complacency), tasks when humans and automation work together (e.g., automation-error costs) and emergent factors that can arise during interactions (e.g., trust in automation). Automation

characteristics such as reliability can provide valuable insight into operator response and performance when an aid performs near perfect or becomes unreliable. For example, if an aid has high reliability, this can lead to misuse, or overreliance on an aid; when an aid has low reliability, this can lead to disuse, or ignoring alerts from an aid or disabling its functions (Parasuraman & Riley, 1997). A study showed that the optimal reliability to be 70% and anything below that point impairs an individual's performance, demonstrating the importance of reliable advice (Wickens & Dixon, 2007). In addition, person factors such as complacency illustrate how individual differences can affect the use of automation. For instance, complacency can occur when automation performance is near perfect resulting in reduced monitoring and vigilance (Parasuraman, Molloy, & Singh, 1993). Previous research on the topic indicates that varying reliability may disrupt complacency (McBride et al., 2014) and complacent behaviors may be due to conditions under multiple-task load (Parasuraman & Manzey, 2010). However, the measurement of complacency and how it is defined is not clearly delineated (Parasuraman et al., 1993). Task variables such as automation-error consequences and accountability can reveal how environmental contexts influence teamwork between humans and automation. For example, accountability in pilot cockpits has been shown to be higher when the accountability is internalized (Mosier et al., 1998) and performance accountability can lead to less automation bias (Skitka, Mosier, & Burdick, 2000). Lastly, emergent factors such as trust in automation or mental workloads are components that can alter how an individual manages errors. A study by Merritt, Heimbaugh, LaChapell, and Lee (2013) investigated trust towards automation with an X-ray luggage-screening task and the

authors concluded that implicit attitudes significantly predicted automation trust.

Furthermore, it has been revealed that appropriately calibrating operator trust can mitigate any potential issues that can arise during human-automation interactions (Lee & See, 2004) and relative trust may be an essential factor involved with a framework for automation use (Dzindolet, Beck, Pierce, & Dawe, 2001). Studies investigating automation use have demonstrated that there are many different variables contributing to how individuals manage errors from automation.

Furthermore, brain activity in response to error monitoring and processing has been measured with fMRI as well as event-related potential (ERP). For instance, a study examining error monitoring during a Go/NoGo task with fMRI and ERP correlations demonstrated that error and conflict monitoring both show involvement with distinct ACC regions (Mathalon, Whitfield, & Ford, 2003). The ACC has been shown to be involved with a wide range of cognitive functions involving decision-making and attention (Bush, Luu, & Posner, 2000), as well as error detection and performance monitoring (Carter et al., 1998; Kiehl, Liddle, & Hopfinger, 2000). Shenhav, Botvinick, and Cohen (2013) postulated that dorsal ACC functionality is based on a model of expected value of control that integrates expected payoffs and rewards. Moreover, cortical activity in sensory brain areas in response to prediction errors has also been examined (Hesselmann, Sadaghiani, Friston, & Kleinschmidt, 2010). A study measuring cortical activity in response to signal detection categories revealed that false alarms evoked more cortical activity than misses, which may be due to individual perceptions involved with each type of error (Ress & Heeger, 2003). There is extensive evidence that

the ACC is involved with error monitoring and that there are perceptual differences involved with each error type; however, the neural basis associated with error processing during unreliable advice from human and machine agents have not been determined and thus warrants further examination.

### **1.3 Overview of the Studies**

The purpose of the studies was to examine how errors moderate advice utilization when comparing humans and machines by revealing the behavioral and neural mechanisms associated with advice taking. Furthermore, the relevance of the research provides insight into the numerous factors that can influence advice utilization by investigating decision-making processes in conjunction with functional magnetic resonance imaging (fMRI) and effective connectivity analysis. Recent behavioral research has provided evidence for advice utilization differences between humans and machines; however, the underlying neural mechanisms involved with human-machine interactions remains to be explored. In both studies, participants partook in an X-ray luggage-screening task where they received good and bad advice from either a machine or human agent framed as an expert, made decisions to accept or reject the advice and then received feedback if their decision was correct or incorrect. Based upon previous findings that false alarms degrade trust and hurt overall performance more than misses (Dixon et al., 2007), we aimed to reveal the influence of bad advice on decision-making processes by manipulating agent reliability with different error types (false alarms, misses). Specifically, in both studies, the reliability of the agents was 60% (40% errors),

but in Chapter Two, the focus was on false alarms while, in Chapter Three, the focus was on misses. We expected performance and advice utilization to be lower in Chapter Two compared to Chapter Three due to the differences in error types. In addition, we expected that errors would decrease overall performance for both studies and that this would ultimately lead to degradation of advice utilization. The differences in advice utilization would be further highlighted when comparing the human agent to the machine agent due to factors such as expectations of reliable advice, agent performance and dispositional credibility associated with each agent. Lastly, we expected that brain regions corresponding with the default-mode network (e.g., TPJ, PreC) and the salience network (e.g., AI, ACC) to be recruited during these studies due to violations of expectations stemming from unreliable advice, salience detection of errors and attribution of dispositional credibility for each agent.

## CHAPTER TWO: THE IMPACT OF FALSE ALARMS ON ADVICE UTILIZATION

### 2.1 Abstract

With new technological advances, advice can come from different sources such as machines or humans, but how individuals respond to such advice and the neural correlates involved need to be better understood. We combined functional MRI and multivariate Granger causality analysis with an X-ray luggage-screening task to investigate the neural basis and corresponding effective connectivity involved with advice utilization from agents framed as experts. Participants were asked to accept or reject good or bad advice from a human or machine agent with manipulated reliability (high false alarm rate). We showed that unreliable advice decreased performance overall and participants interacting with the human agent had a greater depreciation of advice utilization during bad advice. These differences in advice utilization can be due to reevaluation of expectations arising from association of dispositional credibility for each agent. We demonstrated that differences in advice utilization engaged brain regions associated with evaluation of personal characteristics and traits (precuneus, posterior cingulate cortex, temporoparietal junction) and interoception (posterior insula). We found that the right posterior insula and left precuneus were the drivers of the advice utilization network that were reciprocally connected to each other and also projected to all other regions. Our behavioral and neuroimaging results have significant implications

for society because of progressions in technology and increased interactions with machines.

**This work was submitted to *Frontiers in Human Neuroscience*.**

### **Authors**

Kimberly Goodyear, Raja Parasuraman, Sergey Chernyak, Poornima Madhavan, Gopikrishna Deshpande, Frank Krueger

### **Author Contributions**

K.G. and S.C. acquired the data for analysis. K.G., R.P. and F.K. contributed to the conception of the design. K.G., R.P., S.C., P.M., G.D. and F.K. contributed to interpretation of the data. K.G., R.P., S.C., P.M., G.D. and F.K. contributed to drafting of the work and revising it critically. K.G., R.P., S.C., P.M., G.D. and F.K. approved the final version to be published. K.G., R.P., S.C., P.M., G.D. and F.K. agreed to be accountable for all aspects of the work.

### **Funding**

This work was supported by the Air Force of Scientific Research [202857].

## 2.2 Introduction

Individuals often encounter situations in their everyday lives when they must rely on advice from others. With new technological advances, advice can come from not only humans, but also automated devices such as a Global Positioning System. For instance, to provide advanced safety measures, the Transportation Safety Administration (TSA) has implemented X-ray luggage scanners and Advanced Imaging Technology (AIT) for screening passengers and exposing potential security threats (Transportation Safety Administration, 2014). Numerous factors can alter the valuation of advice, such as self-confidence (Bonaccio & Dalal, 2006; Lee & Moray, 1992; Riley, 1996), user trust (P. Madhavan & D. A. Wiegmann, 2007b; Mayer, Davis, & Schoorman, 1995; Rotter, 1967), source credibility (i.e., expert) (Birnbaum & Stegner, 1979; Madhavan & Wiegmann, 2007a; Van Swol & Snizek, 2005) and source reliability/performance (Bonaccio & Dalal, 2006). Understanding how people utilize advice is becoming necessary to provide useful insight for developing safety measures and for appropriate guidelines to predict human behaviors.

Individuals may vary in how they respond to advice and studies have shown that expert advice is more frequently used (Snizek, Schrah, & Dalal, 2004) and more persuasive than novice advice (Jungermann, Fischer, Betsch, & Haberstroh, 2005). In addition, people may respond to advice from automation and humans in similar ways under the premise of a "perfect automation schema," in which an individual believes that automated aids are near perfect (Dzindolet, Pierce, Beck, & Dawe, 2002). Moreover, factors such as dispositional credibility can alter trust between human and machine

advisors due to differences in personal traits such as loyalty or benevolence. For example, it has been postulated that association of dispositional credibility is higher for human agents due to evaluation of personal traits, while automated agents may be judged more by performance levels (Madhavan & Wiegmann, 2007a). However, when expectations of reliable advice are altered due to disconfirmation evidence about an advisor's credibility, decision-making behaviors can be impacted. For example, consistent with disconfirmation theory (Oliver, 1980) decision-making can be affected by initial confirmatory experiences, which can be influenced by bad advice (Staudinger & Buchel, 2013).

Despite existing knowledge of the cognitive processes that affect advice taking, the neural mechanisms and the underlying effective connectivity network involved with good and bad advice from human and machine agents framed as experts remains to be explored. Recent neuroimaging studies have investigated the role of expert advice during decision-making (Boorman, O'Doherty, Adolphs, & Rangel, 2013; Meshi, Biele, Korn, & Heekeren, 2012), social learning (Biele, Rieskamp, Krugel, & Heekeren, 2011; Staudinger & Buchel, 2013) and disobedience (Suen, Brown, Morck, & Silverstone, 2014). Furthermore, the neural activity involved with assigning trait and intentions to others (Mitchell, Macrae, & Banaji, 2006; Saxe & Kanwisher, 2003), self-attributional processes (Cabanis et al., 2013), as well as human-robot interactions during an interactive rock-paper-scissors game (Chaminade et al., 2012) and during observations of social interactions (Wang & Quadflieg, 2015) have been investigated. Overall, key regions associated with the default network (e.g., temporoparietal junction, precuneus, posterior

cingulate cortex, medial prefrontal cortex) and the salience network (dorsal anterior cingulate cortex, bilateral insulae) have been identified in playing a role during advice taking, evaluation of personal traits and during human-robot interactions (Engelmann, Capra, Noussair, & Berns, 2009; Krach et al., 2008).

We aimed to elucidate the neural basis of advice utilization from different agents and the corresponding effective connectivity in the underlying brain network by combining an X-ray luggage-screening task and functional magnetic resonance imaging (fMRI) with multivariate Granger causality analysis. The focus of this study was to examine the impact of false alarms on advice taking behaviors based on previous evidence that false alarms degrade trust and hurt overall performance more than misses (Dixon, Wickens, & McCarley, 2007). On the behavioral level, we hypothesized that unreliable advice would decrease performance (i.e., accuracy) and advice utilization due to disconfirming evidence about the agents' perceived expertise. We further assumed that advice utilization would decrease more during bad advice due to disconfirmation evidence stemming from advice-incongruent experiences (i.e., high false alarm rates) (Dixon et al., 2007) and also over time as errors became more apparent due to participants' reevaluation of the agent's performance (Skitka, Mosier, & Burdick, 2000). In addition, we expected that advice utilization would decrease more for the machine agent compared to the human agent due to differences in dispositional credibility between humans and machines (Madhavan & Wiegmann, 2007a). On the neural level, we first predicted activation differences in brain regions associated with attribution of personal traits and dispositions (Brosch, Schiller, Mojdehbakhsh, Uleman, & Phelps, 2013; Harris,

Todorov, & Fiske, 2005). Secondly, when comparing the human to the machine agent during bad advice over time, we expected regions such as the precuneus and posterior cingulate cortex to be the drivers of the underlying advice utilization network.

## **2.3 Methods**

### **Subjects**

Three studies were conducted according to the ethical guidelines and principles of the Declaration of Helsinki. For the normative rating study, twenty-three male students (age ( $M \pm SD$ ) =  $24.0 \pm 2.6$ ) from George Mason University (GMU) participated to standardize the X-ray luggage images for the experimental studies. For the behavioral study, ten volunteers (6 males, 4 females; age =  $22.3 \pm 2.9$ ) participated to complete an X-ray luggage-screening task without receiving advice. For the fMRI study, twenty-four healthy right-handed volunteers (13 males, 11 females; age =  $20.0 \pm 2.6$ ) determined by the Edinburgh Handedness Inventory (Right-handedness:  $94.5 \pm 7.7$ ) (Oldfield, 1971) participated in the X-ray luggage-screening task while receiving advice either from a human or machine agent. All participants gave written consent approved by GMU's Institutional Review Board and received financial compensation for their participation.

### **Stimuli**

During the normative rating study, the participants rated 320 X-ray images based on three dimensions —clutter ( $4.1 \pm 0.3$ ), general difficulty ( $3.5 \pm 0.4$ ), and confidence in finding the target ( $3.2 \pm 0.6$ )— based on 7-point Likert scales (1 = very low to 7 = very high)

(Madhavan & Gonzalez, 2006). From those images, 64 (32 target and 32 non-target) images were chosen for the experimental studies based on the standardized ratings ([Appendix A.1a](#)).

### **X-ray Luggage-Screening Task**

In the X-ray luggage-screening task, participants were asked to search for the presence or absence of a knife. In the behavioral study, participants did not receive advice and performed the task unassisted; participants in the fMRI study received good (advice-congruent) or bad (advice-incongruent) advice from either a human or machine agent. For both studies, the reliability was set to 60% - good advice: 50% hits (correct alerts) and 10% correct rejections (correct non-alerts); bad advice: 40% false alarms (incorrect alerts) ([Appendix A.1b](#)).

On each trial, the participants saw a set of phases including a fixation cross (0.5 s), advice from one of the agents to “search” or “clear” the bag (2 s), an image of the X-ray luggage (4 s), a decision to accept or reject the advice of the agent to “search” or “clear” the bag (4 s), jitter (~4 s), feedback indicating if their decision was correct or incorrect (2.0 s) and lastly, jitter (~4 s). The jitter times were generated by an fMRI simulator software (<http://www.mccauslandcenter.sc.edu/CRNL/tools/fmrism>) that optimized the timing and consisted of a minimum of 1 seconds and average of 4 seconds ([Appendix A.1c](#)). Participants used response pads to respond and they were given an initial endowment of \$40 and each incorrect answer resulted in a deduction of \$0.30 from the remaining total. Performance, advice utilization, response times, and monetary

deductions were collected during the experiment. The stimuli were presented using E-Prime 2.0 (Psychology Software Tools, Inc., <http://www.pstnet.com/eprime.cfm>).

## **Procedure**

***Pre-Experimental Phase.*** The participants came one to two weeks before the fMRI experiment to complete self-report questionnaires as control measures to investigate individual differences between the agent groups. The control measures included: Interpersonal Reactivity Index (IRI, separate facets of empathy) (Davis, 1983), Complacency-Potential Rating Scale (CPS, feelings towards automation) (Singh, Molloy, & Parasuraman, 1997), National Readiness Technology Scale (NTRS, embracing new technologies) (Parasuraman, 2000), NEO Five-Factor Inventory (NEO-FFI, personality styles) (Costa & McCrae, 1992), and Propensity to Trust (PTT, trust towards automation) (Merritt, Heimbaugh, LaChapell, & Lee, 2013).

***Experimental Phase.*** Before participants completed a practice run for the fMRI experiment, they read descriptions about the human or machine agents (reliability was not disclosed) ([Appendix A.2](#)). They were then asked to rate their trust in and reliability of the human or machine agent on a 10- point Likert scale (0 = very low, 10 = very high). During the four trials of the practice run, participants familiarized themselves with the X-ray luggage-screening task and the five possible knives that could be present in the bags. The participants then completed two runs of the experimental task while in the scanner and afterwards they were again asked to rate reliability and trust.

**Post-Experimental Session.** After the fMRI experiment, participants were asked to rate their confidence in finding the target (i.e., knife) in each of the X-ray luggage images presented during the fMRI experiment on a 10-point Likert scale (1 = very low, 10 = very high).

### **fMRI Data Acquisition**

Imaging data were acquired on a 3 T head-unit only scanner (Siemens Allegra) with a circularly polarized, transmit/receive head coil at the Krasnow Institute for Advanced Study, GMU, Virginia. The anatomical imaging data were based on a 3D T1 weighted MPAGE sequence with TR = 2300 ms, TE = 3.37 ms, flip angle = 7°, slice thickness = 1 mm, voxel dimension = 1 mm x 1 mm x 1 mm and number of slices = 160. The functional imaging data were based on a 2D gradient-echo EPI sequence with TR = 2000 ms, TE = 30 ms, flip angle = 70°, slice thickness = 3 mm, voxel dimensions = 3 mm x 3 mm x 3 mm, number of slices = 33 per volume in an axial orientation parallel to the anterior-posterior commissure. The first two volumes were discarded to allow for T1 equilibrium effects and a total of 330 volumes were taken for each run.

### **Behavioral Data Analysis**

Behavioral data analysis was carried out by Statistical Package for the Social Sciences 20.0 (SPSS 20.0, IBM Corp.) with alpha set to  $p < .05$  (two-tailed). Data were normally distributed (Kolmogorov–Smirnov test) and assumptions for analyses of variance

(Bartlett's test) were not violated. We first investigated task performance (i.e., accuracy) between the agent groups and the no agent group by employing one-way analysis of variance (ANOVA) with Agent (human, machine, no agent) as the between-subjects factor. Next, we looked at advice utilization, response times and monetary deductions with mixed 2 x 2 x 2 repeated-measures ANOVAs with Advice (good, bad) and Time (run 1, run 2) as within-subjects factors and Agent (human, machine) as the between-subjects factor. In addition, we investigated reliability, trust and confidence ratings with mixed 2 x 2 repeated-measures ANOVAs with Time (pre, post) as the within-subjects factor for the reliability/trust ratings and Target (yes, no) as the within-subjects factor for the confidence ratings and with Agent (human, machine) as the between-subjects factor. Lastly, we performed bivariate Spearman's correlations to identify associations between behavioral and control measures as well as independent *t*-tests between the agent groups to investigate group differences.

### **fMRI Data Analysis**

The fMRI data analysis was carried out using NeuroElf software (<http://neuroelf.net>) and BrainVoyager QX 2.8 (Brain Innovation). The functional imaging data were preprocessed using Statistical Parametric Mapping 8 (SPM8, Wellcome Department of Cognitive Neurology) functions batched via NeuroElf, including three-dimensional motion correction (six parameters), slice-scan time correction (temporal interpolation) and a mean functional image was computed for each participant across all runs. The mean functional image was then co-registered with the anatomical images using a joint-

histogram for the different contrast types. Preprocessing of the anatomical images included segmenting images with a unified segmentation procedure (Ashburner & Friston, 2005) and spatial warping were applied to the functional data to normalize the data to a standard Montreal Neurological Institute (MNI) brain template. Lastly, spatial smoothing (Gaussian filter of 6 mm FWHM) was applied to the images to account for any residual differences across participants. A general linear model (GLM) that was corrected for first-order serial correlations was performed (Friston, Harrison, & Penny, 2003). The GLM consisted of thirty-six regressors based on advice utilization (accept, reject) separated by advice (good, bad) and time (run 1, run 2) for each of the five phases (fixation, advice, bag, decision, feedback) on each trial of the X-ray luggage-screening task and six parametric regressors of no interest for the 3D motion correction (translations in X, Y, Z directions, rotations around X, Y, Z axes). The regressor time courses were adjusted for the hemodynamic response delay by convolution with a dual-gamma canonical hemodynamic response function (Buckner, Holmes, Rees, & Friston, 1998). Random-effect analyses were performed at the multi-subject level to explore brain regions associated with the decision and feedback phases.

To reveal brain activations associated with advice utilization, mixed 2 x 2 x 2 ANOVAs on parameter estimates were applied with Advice (good, bad) and Time (run 1, run 2) as within-subjects factors and Agent (human, machine) as the between-subjects factor. For the fMRI results, our main focus was on brain activations during the decision and feedback phases for the three-way interaction since our *a priori* hypotheses was based on the interaction of three factors (advice, time, agent) (see Appendix A.3 for main

effects for the decision and feedback phases). Activations for the decision and feedback phases were reported after correcting for multiple comparisons using a cluster-level statistical threshold (Cluster-level Statistical Threshold Estimator plugin in BrainVoyager QX), which calculates the minimum cluster size to achieve a false activation probability ( $\alpha = 0.05$ ) (Forman et al., 1995; Goebel, Esposito, & Formisano, 2006). The voxel-level threshold was set at  $p < .005$  (uncorrected) and the thresholded map was used for a whole-brain correction criterion based on the estimate of the map's spatial smoothness and on an iterative procedure (Monte Carlo simulation, 1,000 iterations). The activation clusters were displayed in MNI space on an anatomical brain template reversed left to right.

### **Effective Connectivity Analysis**

Investigation of the effective (or directional) brain connectivity in the network of activated brain regions was performed through multivariate Granger causality analysis (GCA) using a custom MATLAB ([www.mathworks.com](http://www.mathworks.com)) code as previously described by Grant et al. (2014), Kapogiannis, Deshpande, Krueger, Thornburg, and Grafman (2014) and Lacey, Stilla, Sreenivasan, Deshpande, and Sathian (2014). Granger causality is based on a temporal precedence concept (Granger, 1969) that can be applied to multivariate effective connectivity modeling of ROI (region of interest) time courses to predict directional influences among brain regions (Deshpande, LaConte, James, Peltier, & Hu, 2009; Friston et al., 2003; Preusse, van der Meer, Deshpande, Krueger, & Wartenburger, 2011; Roebroek, Formisano, & Goebel, 2005; K. Sathian et al., 2011;

Strenziok et al., 2010). The model examines the relationship of variables in time, such that given two variables,  $a$  and  $b$ , if past values of  $a$  better predict the present value of  $b$ , then causality between the variables can be inferred as function of their earlier time points (Hampstead et al., 2011; Krueger, Landgraf, van der Meer, Deshpande, & Hu, 2011; Roebroeck et al., 2005). GCA is advantageous for application of effective connectivity since it is a data-driven approach and there is no requirement for pre-specified connectivity models like dynamic causal modeling (DCM) (Deshpande & Hu, 2012; Deshpande et al., 2009; Deshpande, Sathian, Hu, & Buckhalt, 2012; Friston et al., 2003; Roebroeck et al., 2005). Recent GCA investigations, including experimental applications (Abler et al., 2006) as well as simulations (Deshpande, Sathian, & Hu, 2010b; Wen, Rangarajan, & Ding, 2013), have shown its advantages and validity for assessing effective connectivity.

Based upon on effective connectivity hypotheses, only those regions that survived the fMRI analysis threshold for the interaction effect Advice (good, bad), Time (run 1, run 2), and Agent (human, machine) for the decision phase were selected as ROIs for the subsequent multivariate GCA. Time series of the BOLD (blood-oxygen-level-dependent) signal for the selected ROIs were extracted around peak activation maxima (sphere of  $6 \times 6 \times 6 \text{ mm}^3$ ), averaged across voxels and normalized across participants, per run. Blind hemodynamic deconvolution of the mean ROI BOLD time series was performed using a Cubature Kalman filter, which has been shown to be extremely efficient for jointly estimating latent neural signals and the spatially variable hemodynamic response functions (HRFs) (Havlicek, Friston, Jan, Brazdil, & Calhoun, 2011). In addition, recent

research has shown that this model is not susceptible to over-fitting and produces estimates that are comparable to non-parametric methods (Sreenivasan, Havlicek, & Deshpande, 2015). Hemodynamic deconvolution removes the inter-subject and inter-regional variability of the HRF (Handwerker, Ollinger, & D'Esposito, 2004) as well as its smoothing effect and therefore, increases the effective temporal resolution of the signal. The resulting latent neural signals were entered into a first order dynamic multivariate autoregressive (dMVAR) model for assessing directed interactions between multiple nodes as a function of time (Grant, Wood, Sreenivasan, Wheelock, & White, 2015; Hutcheson et al., 2015; Wheelock et al., 2014)) while factoring out influences mediated indirectly in the set of selected ROIs (Deshpande, Hu, Stilla, & Sathian, 2008; Deshpande, Sathian, & Hu, 2010a; Stilla, Deshpande, LaConte, Hu, & Sathian, 2007). A first order model was implemented because of the interest in causal influences arising from neural delays, which are less than a TR (Deshpande, Libero, Sreenivasan, Deshpande, & Kana, 2013). Furthermore, the dMVAR model's coefficients were allowed to vary as a function of time to obtain condition-specific connectivity values (K Sathian, Deshpande, & Stilla, 2013).

Granger connectivity (GC) path weights for conditions of interest (bad advice) for each agent (human, machine) were extracted. Those corresponding GC path weights were populated into two samples and independent samples *t*-tests were employed to reveal the condition-specific modulations of connectivity ( $q(\text{FDR}) < .05$ ) (Benjamini & Hochberg, 1995), i.e. those paths which had significantly different effective connectivity between human and machine agents while receiving bad advice ([Appendix A.4](#)). Since

GCA is a data-driven approach, the condition-specific modulation was specifically chosen for analysis based upon our fMRI results. Effective connectivity of brain regions (i.e., nodes, edges) was displayed on a brain surface using BrainNet Viewer ([www.nitrc.org/projects/bnv/](http://www.nitrc.org/projects/bnv/)), a graphical interface visualization tool (Xia, Wang, & He, 2013).

## 2.4 Results

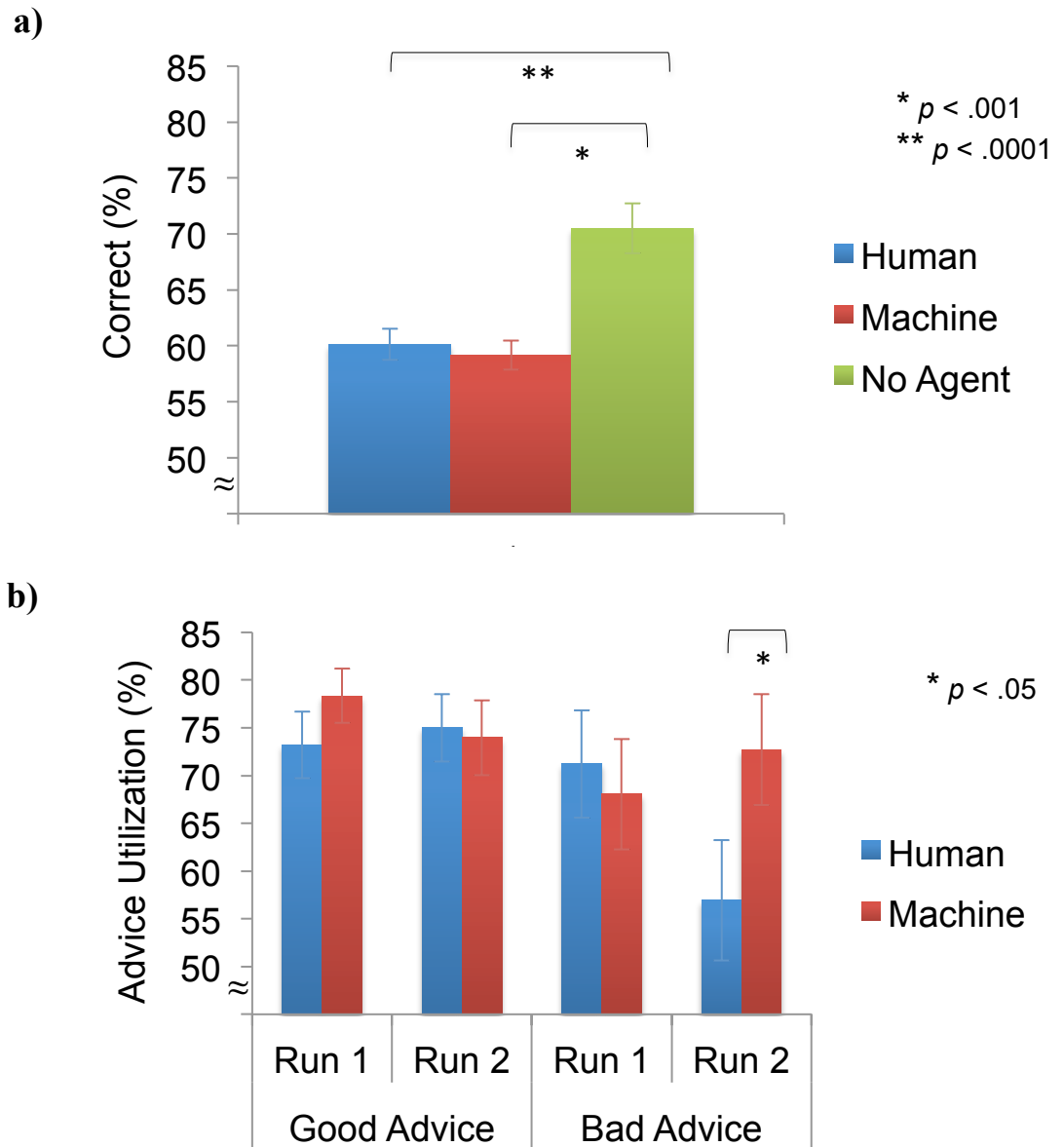
### Behavioral Results

First, we compared the performance between the agent groups and the no advice group by employing a one-way ANOVA with Agent (human, machine, no agent) as between-subjects factors. A significant main effect of Agent ( $F(2, 31) = 13.85, p < .0001$ ) was revealed, and post-hoc testing revealed that the no agent group performed better than the human-agent group ( $t(20) = -4.06, p = .001$ ) and the machine-agent group ( $t(20) = -4.54, p < .0001$ ) (**Fig. 1a**). Next, we looked at advice utilization, response times, and monetary deductions. For *advice utilization*, a significant main effect of Advice was revealed ( $F(1,22) = 7.63, p = .011$ ), indicating that participants accepted good advice more than bad advice. In addition, a significant three-way interaction of Advice x Time x Agent was identified ( $F(1, 22) = 5.06, p = .035$ ), but no significant main effects of Agent ( $F(1, 22) = 0.65, p = .429$ ) or Time ( $F(1, 22) = 2.30, p = .144$ ) and no significant two-way interaction effects of Advice x Agent ( $F(1, 22) = 0.56, p = .463$ ), Time x Agent ( $F(1, 22) = 2.54, p = .125$ ), and Advice x Time ( $F(1, 22) = 0.40, p = .536$ ) (**Fig. 1b**) were found. Follow-up 2 x 2 ANOVAs showed a significant interaction effect of Time x Agent for

bad advice ( $F(1, 22) = 5.63, p = .027$ ), but not for good advice ( $F(1, 22) = 1.23, p = .279$ ). Follow-up independent samples  $t$ -tests revealed that the human-agent group accepted bad advice less than the machine-agent group during run 2 ( $t(22) = -1.84, p = .040$ ).

For *response times*, significant main effects of Advice ( $F(1, 22) = 12.26, p = .002$ ) and Time ( $F(1, 22) = 5.85, p = .024$ ) were found, indicating that responses were faster during good compared to bad advice and during run 2 compared to run 1 ([Appendix A.5a](#)). A marginally significant interaction effect was found for the interaction of Time x Agent ( $F(1, 22) = 4.35, p = .049$ ), but no significant main effect of Agent ( $F(1, 22) = 0.49, p = .491$ ) and no significant interaction effects of Advice x Agent ( $F(1, 22) = 0.10, p = .758$ ), Advice x Time ( $F(1, 22) = 0.07, p = .798$ ), and Advice x Time x Agent ( $F(1, 22) = 0.06, p = .811$ ) were found.

For *monetary deductions*, a significant main effect of Advice ( $F(1, 22) = 292.45, p < .0001$ ) was revealed, indicating that deductions were higher during bad advice compared to good advice ([Appendix A.5b](#)). In addition, a marginally significant interaction effect of Time x Agent was found ( $F(1, 22) = 4.61, p = .043$ ), but no significant main effects of Time ( $F(1, 22) = 0.31, p = .583$ ) and Agent ( $F(1, 22) = 1.56, p = .224$ ), or interaction effects of Advice x Agent ( $F(1, 22) = 0.10, p = .758$ ), Advice x Time ( $F(1, 22) = 0.10, p = .921$ ), and Advice x Time x Agent ( $F(1, 22) = 0.09, p = .768$ ) were found.

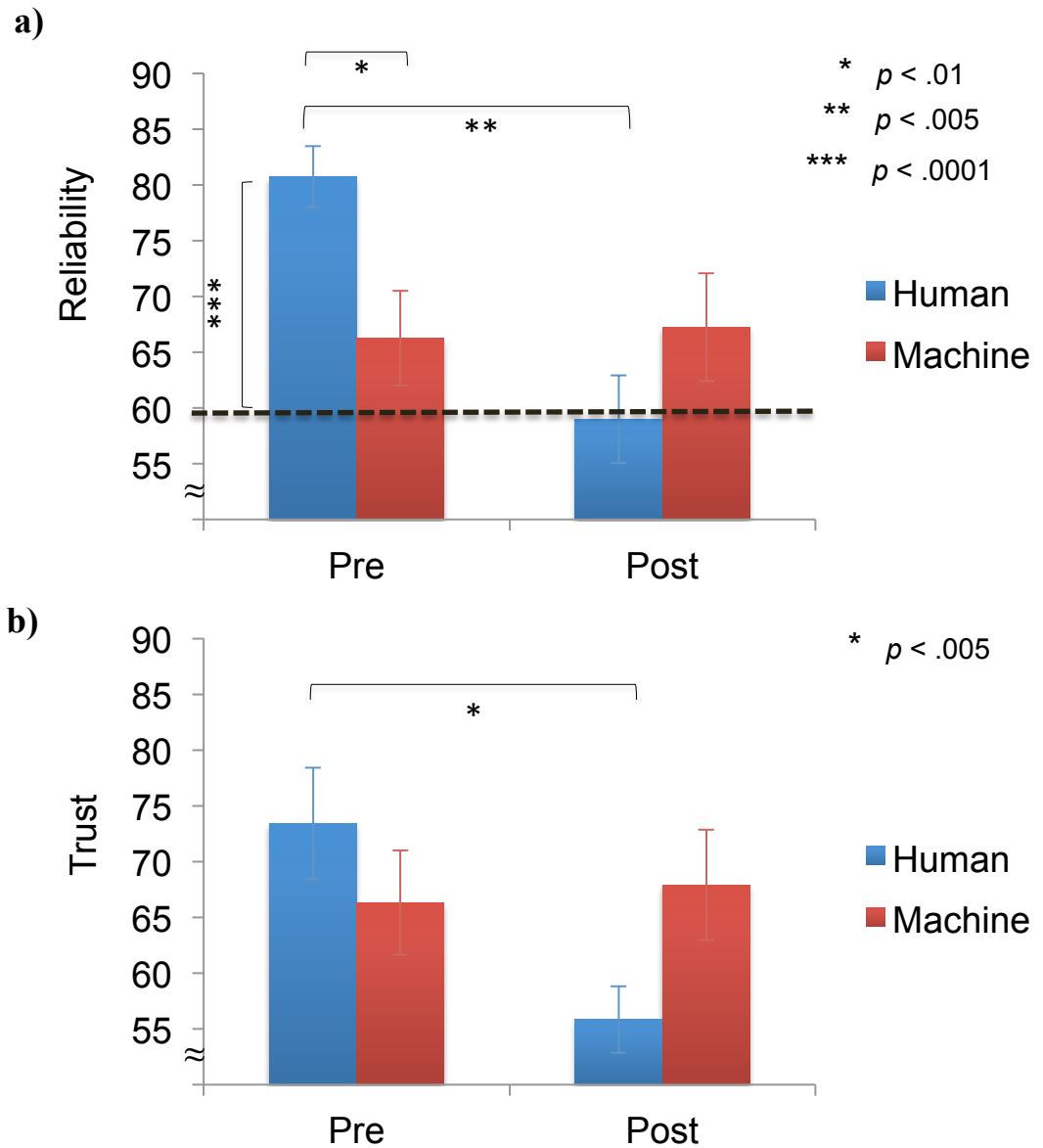


**Figure 1. False Alarm Behavioral Results**

**Results for the Decision Phase ( $M \pm SEM$ ).** **a) Task Performance.** The no agent group performed better than human- and machine-agent groups. **b) Advice Utilization.** Advice utilization during bad advice from the human agent was significantly lower during run 2 compared to the machine agent.

In addition, we looked at pre- and post-experiment ratings (reliability, trust) using repeated-measures ANOVAs with Time (run 1, run 2) and Agent (human, machine) as factors. The *reliability ratings* showed no significant main effect of Agent ( $F(1, 22) = 0.62, p = .439$ ), but a significant main effect of Time ( $F(1, 22) = 6.54, p = .018$ ) and a significant interaction effect of Time x Agent ( $F(1, 22) = 7.86, p = .010$ ) (Fig. 2.2a). Post-hoc testing revealed that the human agent's pre-reliability was rated higher than the machine's pre-reliability ( $t(22) = 2.87, p = .009$ ) and the human's reliability ratings decreased from pre- to post-experiment ( $t(11) = 4.10, p = .002$ ). Furthermore, one-sample *t*-tests on perceived versus actual reliability (60%) of the agent showed that pre-reliability ratings were significantly higher than the actual reliability for the human agent ( $t(11) = 6.79, p < .0001$ ).

For *trust ratings*, no significant main effects of Agent ( $F(1, 22) = 0.26, p = .615$ ) and Time ( $F(1, 22) = 3.96, p = .059$ ) were observed, but a significant interaction effect of Time x Agent ( $F(1, 22) = 5.89, p = .026$ ) was demonstrated (Fig. 2.2b). Post-hoc testing revealed that trust ratings significantly decreased from pre- to post-experiment for the human agent ( $t(11) = 4.18, p = .002$ ). For *confidence ratings*, no main effect of Agent ( $F(1, 22) = 4.16, p = .054$ ) or significant interaction effect of Target x Agent ( $F(1, 22) = 2.46, p = .131$ ) were found, but a significant main effect of Target ( $F(1, 22) = 53.44, p < .0001$ ) was revealed, indicating that confidence was rated higher on target bags compared to non-target bags. (Appendix A.6).



### Figure 2. False Alarm Rating Results

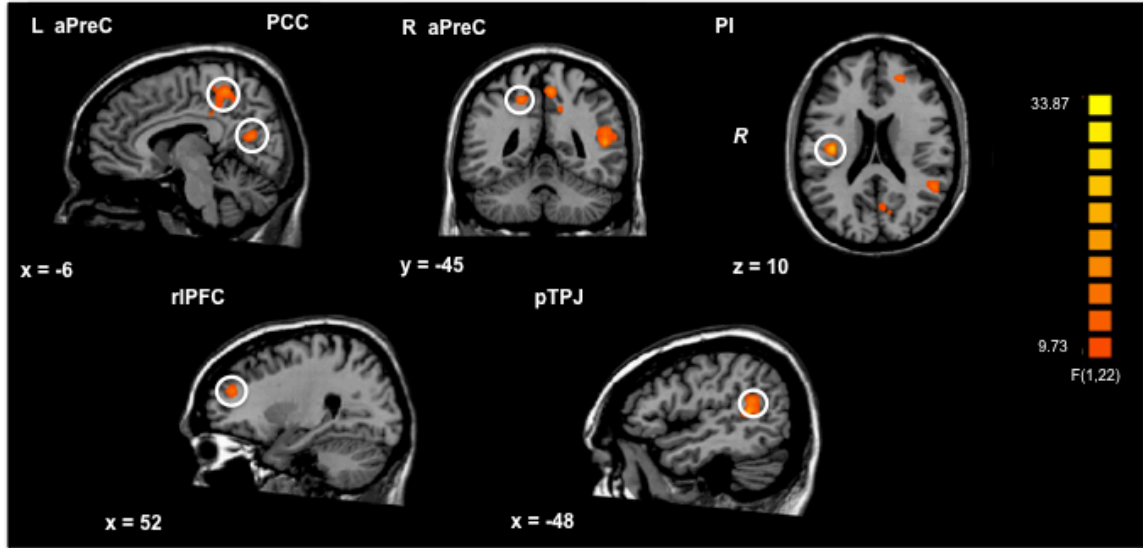
**Results For Ratings ( $M \pm SEM$ ).** **a) Pre- And Post-Reliability.** Pre-reliability was higher for the human agent compared to the machine agent. For the human agent, perceived pre-reliability was significantly higher than the actually reliability of the agent (60%) and post-reliability ratings significantly decreased. **b) Pre- And Post-Trust.** Post-trust was significantly lower than pre-trust for the human agent.

Finally, we analyzed at differences in control measures (e.g., demographic measures and questionnaires) with bivariate Spearman's  $\rho$  correlations and independent samples  $t$ -tests. For the human-agent group, a positive correlation between the NTRS insecurity score and pre-reliability ratings ( $r(12) = .738, p = .006$ ) and pre-trust ratings ( $r(12) = .733, p = .007$ ) were found, indicating that a higher insecurity score towards automation (i.e., greater preference towards human interactions) was positively associated with higher pre-reliability and pre-trust ratings. No significant group differences were identified for any of the control measures ([Appendix A.7](#)).

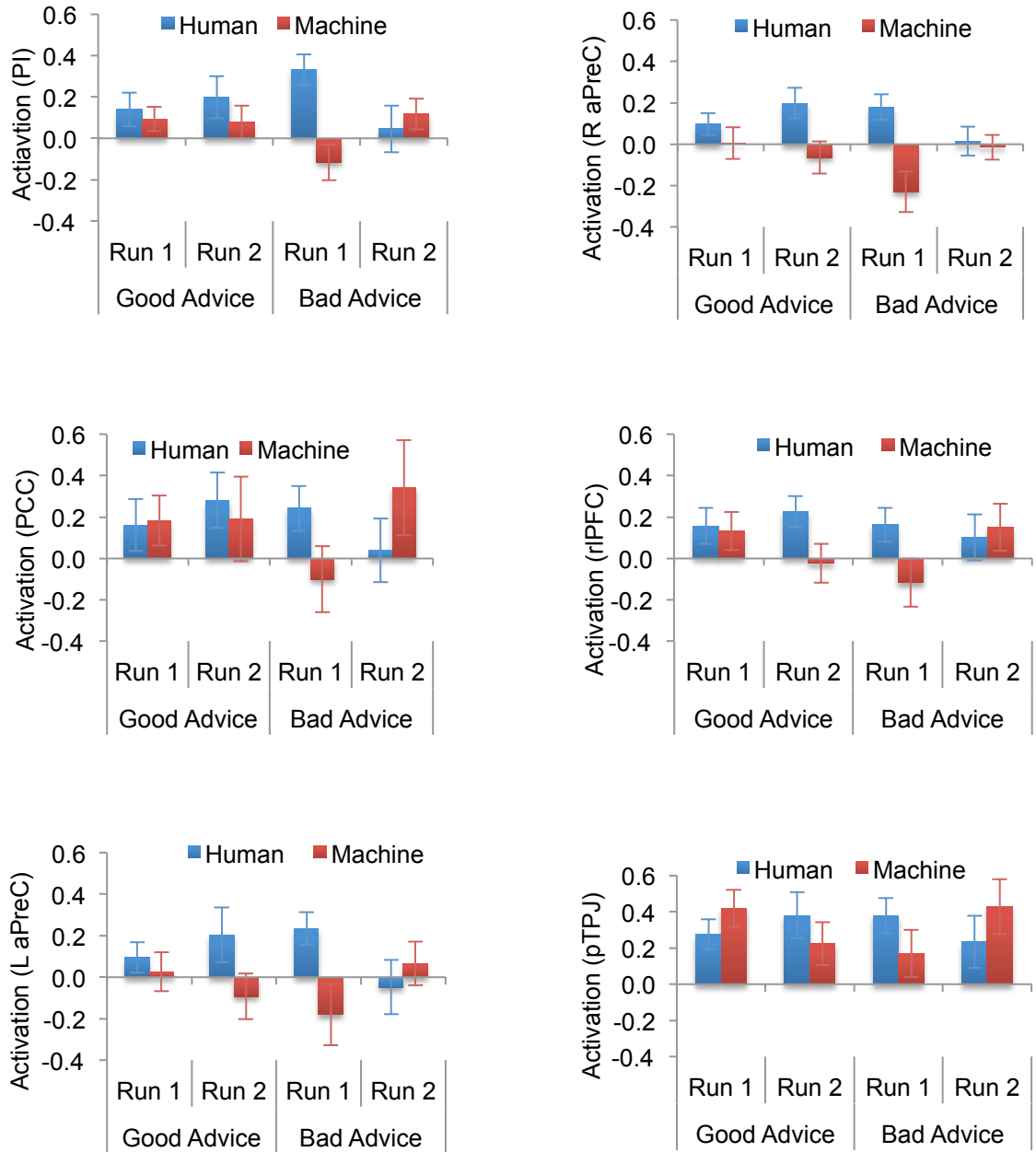
### Neuroimaging Results

For the fMRI results, we looked at brain activations during the decision and feedback phases for the three-way interaction. For the *decision phase*, a significant three-way interaction effect ( $\alpha < .05, k = 21$ ) was found in the right (R) posterior insula (PI) (BA 13); R anterior precuneus (aPreC) (BA 5/7), left (L) aPreC (BA 5/7); L posterior cingulate cortex (PCC) (BA 30/31); L rostrolateral prefrontal cortex (rlPFC) (superior frontal gyrus: SFG; BA 10); and L posterior temporoparietal junction (pTPJ) (superior temporal gyrus: STG; BA 22) ([Fig. 3, Fig. 4, Tab. 1](#)). The results indicate that there was higher activation during run 1 for the human-agent group compared to machine-agent group during bad advice. For the *feedback phase*, a significant three-way interaction ( $\alpha < .05, k = 14$ ) was found in the L dorsomedial prefrontal cortex (dmPFC) (medial frontal gyrus: MFG; BA 9/10) showing higher activation for the human agent during run 2 for good compared to bad advice ([Fig. 5, Tab. 1](#)). Note that no further post-hoc comparisons

were performed on the extracted data from the decision or feedback phases to avoid non-independent analyses, or double dipping (Kriegeskorte, Simmons, Bellgowan, & Baker, 2009).

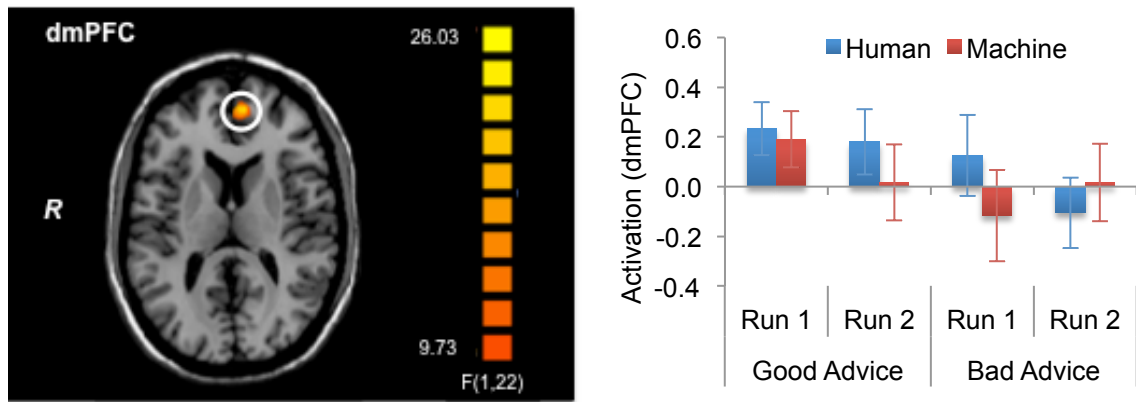


**Figure 3. False Alarm Brain Activations for Decision Phase** ( $\alpha < .05$ ,  $k = 21$ ). The three-way interaction (Advice x Run x Agent) during the decision phase significantly activated the right posterior insula (PI), right anterior precuneus (aPreC), left aPreC, left posterior cingulate cortex (PCC), left rostrolateral prefrontal cortex (rlPFC) and left posterior temporoparietal junction (pTPJ).



**Figure 4. False Alarm Activation Patterns During Decision Phase**

The activation pattern indicates higher activation for the human- compared to machine-agent group for bad advice during run 1. The bar plots shown are for visualization purposes.



**Figure 5. False Alarm Brain Activations During Feedback Phase** ( $\alpha < .05$ ,  $k = 14$ ). The three-way interaction (Advice x Run x Agent) during the feedback phase significantly activated the left dorsomedial prefrontal cortex (dmPFC). The activation pattern shows lower activation for bad advice compared to good advice during run 2 for the human agent. The bar plot serves as a visual aid for the activation pattern.

**Table 1. False Alarm Brain Regions**

**Brain Regions Associated with the Three-Way Interaction.** Brain regions showing significant activation clusters associated during the decision (minimum cluster of 21) and feedback (minimum cluster of 14) phases ( $\alpha < .05$ , cluster-level threshold corrected). PI, posterior insula (BA 13); aPreC, anterior precuneus (BA 5/7); PCC, posterior cingulate cortex (BA 30/31); rLPFC, rostrolateral prefrontal cortex (BA 10); pTPJ, posterior temporoparietal junction BA 22); dmPFC, dorsomedial prefrontal cortex (BA 9/10).

	<i>F</i> (1,22) value	Cluster Size (mm <sup>3</sup> )	x	y	z
<b>Decision phase</b>					
<i>(Advice x Run x Agent)</i>					
Right posterior insula	32.86	854	36	-15	21
Right anterior precuneus	18.65	593	18	-42	45
Left anterior precuneus	21.52	2214	-6	-42	51
Left posterior cingulate cortex	24.96	607	-3	-63	15
Left rostrolateral prefrontal cortex	17.34	692	-21	45	21
Left posterior temporoparietal junction	23.58	1678	-48	-45	9
<b>Feedback phase</b>					
<i>(Advice x Run x Agent)</i>					
Left dorsomedial prefrontal cortex	25.03	655	-6	51	12

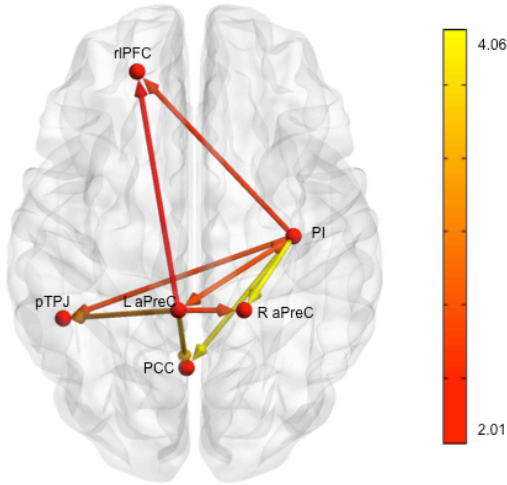
## Effective Connectivity Results

Based on our fMRI results, we implemented multivariate GCA to identify effective connectivity among brain regions during the decision phase when comparing the human with the machine agent during bad advice for run 1 (all connections survived  $q(\text{FDR}) < .05$ , except the connections to the L rLPFC that survived  $q(\text{FDR}) < .08$ ) (Tab. 2).

Analysis for the feedback phase was not done due to the fact that only one region survived for the fMRI results. The L aPreC and PI were identified as the source ROIs; they were the drivers of the network making reciprocal connections to each other, while

also both sending output connections to all target ROIs (R aPreC, PCC, rIPFC and pTPJ)

(Fig. 6).



### Figure 6. False Alarm Results for Multivariate Granger Causality Analysis

The effective connectivity network for bad advice during the decision phase for run 1 when comparing the human with machine agent showed that the PI (posterior insula) and L aPreC (anterior precuneus) were drivers of the network and also the source ROIs for all other target ROIs (R aPreC, PCC (posterior cingulate cortex), rIPFC (rostrolateral prefrontal cortex) and pTPJ (posterior temporoparietal junction). Note that all connections survived  $q(\text{FDR}) < .05$ , except the connections to rIPFC that survived  $q(\text{FDR}) < .08$ . The color bar represents the t-value of the comparisons shown in Table 2.

**Table 2. False Alarm Granger Causality Analysis**

**Path Weights for Granger Causality Analysis.** The path weights displayed show significant effective connectivity paths that are stronger in the human-agent group compared to the machine-agent group during run 1 (all connections survived  $q(\text{FDR}) < .05$ , except the connection to rLPFC that survived  $q(\text{FDR}) < .08$ ). The directionality of the connectivity is shown in the first two columns, with the source column showing the ROIs that predict activation in the target column ROIs. The strength of connectivity is given by the mean path weights in the third column. PI, posterior insula; aPreC, anterior precuneus; PCC, posterior cingulate cortex; rLPFC, rostrolateral prefrontal cortex; pTPJ, posterior temporoparietal junction.

Source	Target	Path weight		<i>t</i> value	<i>p</i> value
		Human	Machine		
PI	R aPreC	0.23	0.18	4.06	$2.80 \times 10^{-5}$
	L aPreC	0.18	0.19	2.57	$5.16 \times 10^{-3}$
	PCC	0.27	0.18	3.96	$4.16 \times 10^{-5}$
	rLPFC	0.16	0.18	2.32	$1.04 \times 10^{-2}$
	pTPJ	0.17	0.15	2.52	$6.02 \times 10^{-3}$
L aPreC	PI	0.18	-0.17	2.42	$7.80 \times 10^{-3}$
	R aPreC	0.18	-0.12	2.44	$7.51 \times 10^{-3}$
	PCC	0.20	-0.15	3.47	$2.79 \times 10^{-4}$
	rLPFC	0.16	-0.15	2.01	$2.22 \times 10^{-2}$
	pTPJ	0.24	-0.21	3.12	$9.39 \times 10^{-4}$

## 2.5 Discussion

The purpose of this research was to understand the neural basis and corresponding effective connectivity network involved during advice utilization from human and machine agents framed as experts. To provide a greater understanding of the behavioral and neural underpinnings associated with advice taking, we manipulated agent reliability with a high false alarm rate to reveal the decision-making processes during good and bad

advice. We first revealed that unreliable advice decreased performance, which has been previously reported by other behavioral studies investigating advice differences between humans and machines (Dzindolet et al., 2002; Madhavan & Wiegmann, 2007a). An earlier study investigating credibility found that advice utilization decreased for expert automation but not for expert humans; however, this study focused entirely on misses and false alarms, which could account for any differences between these earlier findings and ours (Madhavan & Wiegmann, 2007a). In addition, a study investigating perception during a contrast-detection task showed that false alarms evoked more cortical activity when compared to misses, which supports the notion that participants' percepts may vary when presented with different types of errors (Ress & Heeger, 2003). In our study, we focused only on false alarms since there is evidence of distinct neuronal activity associated with false alarms when compared to misses and behavioral studies have demonstrated differences between the two error types (Dixon et al., 2007; McBride, Rogers, & Fisk, 2014)

Contradictory to our hypothesis, the behavioral results revealed that the decline in advice utilization was greater for the human agent compared to the machine agent. We expected that advice utilization would degrade faster for the machine agent because of differences in association of dispositional credibility; however, our results indicate that false alarms weighed more heavily on the human-agent group. Our findings provide evidence that although assignment of personal traits may have been higher for the human agent, the prevalence of false alarms may have altered evaluations of performance levels due to the type of error presented. Furthermore, to reveal any preconceived notions that

participants had about the human and machine agents, we examined whether the perceived pre-reliability differed from the actual reliability for each agent. Interestingly, the human agent's pre-reliability was rated significantly higher than the actual reliability, showing that the human-agent group expected their advisor to be more reliable. Our finding supports other behavioral studies that indicate that preconceived notions can influence participants' perceptions of advice (Madhavan & Wiegmann, 2007b). Pre-reliability and pre-trust ratings for the human agent showed a positive association with insecurity scores for embracing new technologies, indicating that participants interacting with the human agent had initial inclinations that tended towards human interactions. These findings indicate that participants interacting with the human agent could have perceivably built a mental model of their expectations about the agent's credibility and deviations from expected behavior likely caused a reevaluation of the human agent's performance (Burgoon, 1993). The change in perspectives would ultimately cause a shift towards self-reliance and possibly increased responsibility/accountability for the outcome of their decisions (Dzindolet et al., 2002). Post-reliability ratings for the human-agent group showed a shift towards the actual reliability of the agent, which indicates that the human-agent group was able to discern the agent's performance and recalibrate their expectations. Moreover, post-trust was lower than pre-trust for human agent, supporting previous evidence that false alarms degrade trust (Dixon et al., 2007; Rice & McCarley, 2011). Lastly, our results cannot be explained by any of our control measures or confidence ratings because we found no differences between the agent groups.

Moreover, our results revealed that advice utilization decreased during bad advice compared to good advice. Since bad advice was advice-incongruent, it could have created a mismatch between what the participants perceived and what they were advised, resulting in disconfirmation experiences. The discrepancies during advice-disconfirmation experiences most likely lead to skepticism during bad advice and ultimately degradation of advice utilization. As a consequence, response times for both groups were slower during bad advice, since participants had more conflicting perceptual processes (advice-incongruencies). In addition, monetary deductions were higher overall for bad advice, indicating that bad advice caused participants to make more erroneous decisions.

Subsequently, we identified the neural basis and effective connectivity of the underlying brain network associated with advice utilization. On the neural level, we had two expectations regarding brain activity. First, we expected activation differences in regions associated with attribution of personal traits and dispositions, (Brosch et al., 2013; Harris et al., 2005), and secondly, when comparing the agent groups during bad advice over time, brain regions such as the precuneus and posterior cingulate cortex would be the drivers of the advice utilization network. Our neuroimaging results revealed brain regions associated with domain-general large-scale networks, such as the default-mode network (left pTPJ, bilateral aPreC, left PCC) typically engaged in social evaluations, the salience network (PI) for detection of internal and external salient events, and the central-executive network (left rLPFC) implicated in higher-order executive functions (Menon, 2011). Similarly to our fMRI hypotheses, on the effective

connectivity level, we theorized that a network to be differentially involved when comparing the human to the machine agent for bad advice during run 1. Our effective connectivity analysis revealed that left aPreC and PI were drivers of the network that were reciprocally connected to each other. The aPreC and PI acted as centralized hubs of the network, presumably by integrating social evaluations (e.g., judgments about other's intentions and personal traits) (Cavanna & Trimble, 2006) with interoception (e.g., recruitment of physiological responses to environmental cues) (Kurth et al., 2010). Previous evidence supports the notion that integration of subjective mental states (PreC) and information about internal bodily states (anterior insula, AI) are important for awareness of one's emotional state (Terasawa, Fukushima, & Umeda, 2013). Since participants interacting with the human agent could have had greater conceptualization of the discrepancies between the actual and perceived reliability, this could have led to a visceral response (PI) to the unreliable advice in conjunction with association of personal traits (aPreC) during interactions with the agent.

Furthermore, our effective connectivity results indicated that both hubs (left aPreC, PI) had directional influences on all other regions (right aPreC, left pTPJ, PCC, and left rLPFC) to guide decision-making processes during advice utilization. PreC activation has been identified during a comparison of other- versus self-attribution, showing the involvement of this region during causal attributions towards another (Farrer & Frith, 2002). In addition, PCC activation has been implicated in adapting behaviors (Pearson, Heilbronner, Barack, Hayden, & Platt, 2011) and self-reflection (Johnson et al., 2002), while the pTPJ has been shown to be activated during social cognitions such as

determining intentionality of others (Mars et al., 2012). Other fMRI studies investigating expert advice have shown activation in PCC and PreC during no advice conditions (Engelmann et al., 2009) and in regions such as PCC, insula and medial frontal gyrus when comparing advice vs. no advice in experts and peers (Suen et al., 2014); however, we did not expect equivalent results since our experimental design looked at differences between humans and machines. Furthermore, we found directional influences to the rIPFC, which is part of the central-executive network and has shown to be involved in reasoning (Christoff et al., 2001) and while making uncertain decisions (Badre, Doll, Long, & Frank, 2012).

In addition to our results for the decision phase, we also expected participants to have a heightened awareness of bad advice due to feedback, which would ultimately lead to a behavioral adjustment in advice utilization over time. During the feedback phase, we found activation in the dmPFC, which coincides with another study that showed dmPFC activity during feedback after iterative trials with the same advisor (Behrens, Hunt, Woolrich, & Rushworth, 2008). The dmPFC has been shown to be involved with social cognition (Amodio & Frith, 2006) and during inferences about other's goals and traits (Krueger, Grafman, & McCabe, 2008; Van Overwalle, 2009). In our study, participants interacting with the human agent showed lower dmPFC activation during bad compared to good advice toward the end of the experiment, which shows that, as participants ascertained that the human agent was unreliable, they could have placed lower value on bad advice while receiving feedback.

Our study had a few limitations that should be addressed. First, we looked at differences between good and bad advice by manipulating agent reliability with only false alarms. Future studies could elaborate on our findings by investigating how misses degrade advice utilization between humans and machines and the effective connectivity network associated with those differences. Furthermore, to prevent cognitive anchoring, or the tendency to rely too heavily on the first piece of information acquired, we had participants receive advice before they made their decisions, rather than receiving advice after they made their decisions. Cognitive anchoring has been shown to decrease reliance on automated aids during self-generated decisions (Madhavan & Wiegmann, 2005) and future studies could investigate this phenomena by implementing a paradigm where participants receive advice after they make their decisions.

In summary, our findings provide extensive insight into underlying factors involved with advice utilization from humans and machines and the differences that account for those behaviors. Our results have significant implications for society because of progressions in technology and increased interactions with machines. A greater discernment of the various facets involved with machine interactions will ultimately serve to calibrate behavioral responses and to optimize future safety guidelines. Understanding the variables and environmental differences involved during advice taking will allow for substantive information to improve security and ultimately prevent potential catastrophic disasters.

## 2.6 References

- Abler, B., Roebroek, A., Goebel, R., Höse, A., Schönfeldt-Lecuona, C., Hole, G., & Walter, H. (2006). Investigating directed influences between activated brain areas in a motor-response task using fMRI. *Magnetic Resonance Imaging*, 24(2), 181-185.
- Amodio, D. M., & Frith, C. D. (2006). Meeting of minds: the medial frontal cortex and social cognition. *Nat Rev Neurosci*, 7(4), 268-277.
- Ashburner, J., & Friston, K. J. (2005). Unified segmentation. *Neuroimage*, 26(3), 839-851.
- Badre, D., Doll, B. B., Long, N. M., & Frank, M. J. (2012). Rostrolateral prefrontal cortex and individual differences in uncertainty-driven exploration. *Neuron*, 73(3), 595-607.
- Behrens, T. E. J., Hunt, L. T., Woolrich, M. W., & Rushworth, M. F. S. (2008). Associative learning of social value. *Nature*, 456(7219), 245-249.
- Benjamini, Y., & Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1), 289-300.
- Biele, G., Rieskamp, J., Krugel, L. K., & Heekeren, H. R. (2011). The Neural Basis of Following Advice. *PLoS Biol*, 9(6), e1001089.
- Birnbaum, M. H., & Stegner, S. E. (1979). Source credibility in social judgment: Bias, expertise, and the judge's point of view. *Journal of Personality and Social Psychology*, 37(1), 48.
- Bonaccio, S., & Dalal, R. S. (2006). Advice taking and decision-making: An integrative literature review, and implications for the organizational sciences. *Organizational Behavior and Human Decision Processes*, 101(2), 127-151.
- Boorman, E. D., O'Doherty, J. P., Adolphs, R., & Rangel, A. (2013). The behavioral and neural mechanisms underlying the tracking of expertise. *Neuron*, 80(6), 1558-1571.
- Brosch, T., Schiller, D., Mojdehbakhsh, R., Uleman, J. S., & Phelps, E. A. (2013). Neural mechanisms underlying the integration of situational information into attribution outcomes. *Soc Cogn Affect Neurosci*, 8(6), 640-646.
- Buchel, C., Holmes, A. P., Rees, G., & Friston, K. J. (1998). Characterizing Stimulus-Response Functions Using Nonlinear Regressors in Parametric fMRI Experiments. *Neuroimage*, 8, 140-148.
- Burgoon, J. K. (1993). Interpersonal Expectations, Expectancy Violations, and Emotional Communication. *Journal of Language and Social Psychology*, 12(1-2), 30-48.
- Cabanis, M., Pyka, M., Mehl, S., Muller, B. W., Loos-Jankowiak, S., Winterer, G., . . . Kircher, T. (2013). The precuneus and the insula in self-attributional processes. *Cogn Affect Behav Neurosci*, 13(2), 330-345.
- Cavanna, A. E., & Trimble, M. R. (2006). The precuneus: a review of its functional anatomy and behavioural correlates. *Brain*, 129(Pt 3), 564-583.

- Chaminade, T., Rosset, D., Da Fonseca, D., Nazarian, B., Lutchter, E., Cheng, G., & Deruelle, C. (2012). How do we think machines think? An fMRI study of alleged competition with an artificial intelligence. *Frontiers in human neuroscience*, 6.
- Christoff, K., Prabhakaran, V., Dorfman, J., Zhao, Z., Kroger, J. K., Holyoak, K. J., & Gabrieli, J. D. (2001). Rostrolateral prefrontal cortex involvement in relational integration during reasoning. *Neuroimage*, 14(5), 1136-1149.
- Costa, P., & McCrae, R. (1992). *Revised NEO Personality Inventory (NEO PI-R) and NEO Five-Factor Inventory (NEO-FFI) professional manual*. Odessa, FL: Psychological Assessment Resources, Inc.
- Davis, M. H. (1983). Measuring individual differences in empathy: Evidence for a multidimensional approach. *Journal of Personality and Social Psychology*, 44(1), 113-126.
- Deshpande, G., & Hu, X. (2012). Investigating effective brain connectivity from fMRI data: past findings and current issues with reference to Granger causality analysis. *Brain Connect*, 2(5), 235-245.
- Deshpande, G., Hu, X., Stilla, R., & Sathian, K. (2008). Effective connectivity during haptic perception: A study using Granger causality analysis of functional magnetic resonance imaging data. *Neuroimage*, 40(4), 1807-1814.
- Deshpande, G., LaConte, S., James, G. A., Peltier, S., & Hu, X. (2009). Multivariate Granger causality analysis of fMRI data. *Hum Brain Mapp*, 30(4), 1361-1373.
- Deshpande, G., Libero, L. E., Sreenivasan, K. R., Deshpande, H. D., & Kana, R. K. (2013). Identification of neural connectivity signatures of autism using machine learning. *Front Hum Neurosci*, 7, 670.
- Deshpande, G., Sathian, K., & Hu, X. (2010a). Assessing and compensating for zero-lag correlation effects in time-lagged Granger causality analysis of FMRI. *IEEE Trans Biomed Eng*, 57(6), 1446-1456.
- Deshpande, G., Sathian, K., & Hu, X. (2010b). Effect of hemodynamic variability on Granger causality analysis of fMRI. *Neuroimage*, 52(3), 884-896.
- Deshpande, G., Sathian, K., Hu, X., & Buckhalt, J. A. (2012). A rigorous approach for testing the constructionist hypotheses of brain function. *Behav Brain Sci*, 35(3), 148-149.
- Dixon, S. R., Wickens, C. D., & McCarley, J. S. (2007). On the independence of compliance and reliance: Are automation false alarms worse than misses? *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 49(4), 564-572.
- Dzindolet, M. T., Pierce, L. G., Beck, H. P., & Dawe, L. A. (2002). The perceived utility of human and automated aids in a visual detection task. *Human Factors*, 44(1), 79-94.
- Engelmann, J. B., Capra, C. M., Noussair, C., & Berns, G. S. (2009). Expert Financial Advice Neurobiologically “Offloads” Financial Decision-Making under Risk. *PloS one*, 4(3), e4957.
- Farrer, C., & Frith, C. D. (2002). Experiencing oneself vs another person as being the cause of an action: the neural correlates of the experience of agency. *Neuroimage*, 15(3), 596-603.

- Forman, S. D., Cohen, J. D., Fitzgerald, M., Eddy, W. F., Mintun, M. A., & Noll, D. C. (1995). Improved assessment of significant activation in functional magnetic resonance imaging (fMRI): use of a cluster-size threshold. *Magn Reson Med*, 33(5), 636-647.
- Friston, K. J., Harrison, L., & Penny, W. (2003). Dynamic causal modelling. *Neuroimage*, 19(4), 1273-1302.
- Goebel, R., Esposito, F., & Formisano, E. (2006). Analysis of functional image analysis contest (FIAC) data with brainvoyager QX: From single-subject to cortically aligned group general linear model analysis and self-organizing group independent component analysis. *Hum Brain Mapp*, 27(5), 392-401.
- Granger, C. W. (1969). Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: Journal of the Econometric Society*, 424-438.
- Grant, M. M., White, D., Hadley, J., Hutcheson, N., Shelton, R., Sreenivasan, K., & Deshpande, G. (2014). Early life trauma and directional brain connectivity within major depression. *Hum Brain Mapp*, 35(9), 4815-4826.
- Grant, M. M., Wood, K., Sreenivasan, K., Wheelock, M., & White, D. (2015). Influence of Early Life Stress on Intra- and Extra-Amygdaloid Causal Connectivity.
- Hampstead, B. M., Stringer, A. Y., Stilla, R. F., Deshpande, G., Hu, X., Moore, A. B., & Sathian, K. (2011). Activation and effective connectivity changes following explicit-memory training for face-name pairs in patients with mild cognitive impairment: a pilot study. *Neurorehabil Neural Repair*, 25(3), 210-222.
- Handwerker, D. A., Ollinger, J. M., & D'Esposito, M. (2004). Variation of BOLD hemodynamic responses across subjects and brain regions and their effects on statistical analyses. *Neuroimage*, 21(4), 1639-1651.
- Harris, L. T., Todorov, A., & Fiske, S. T. (2005). Attributions on the brain: neuro-imaging dispositional inferences, beyond theory of mind. *Neuroimage*, 28(4), 763-769.
- Havlicek, M., Friston, K. J., Jan, J., Brazdil, M., & Calhoun, V. D. (2011). Dynamic modeling of neuronal responses in fMRI using cubature Kalman filtering. *Neuroimage*, 56(4), 2109-2128.
- Hutcheson, N. L., Sreenivasan, K. R., Deshpande, G., Reid, M. A., Hadley, J., White, D. M., . . . Lahti, A. C. (2015). Effective connectivity during episodic memory retrieval in schizophrenia participants before and after antipsychotic medication. *Hum Brain Mapp*, 36(4), 1442-1457.
- Johnson, S. C., Baxter, L. C., Wilder, L. S., Pipe, J. G., Heiserman, J. E., & Prigatano, G. P. (2002). *Neural correlates of self - reflection* (Vol. 125).
- Jungermann, H., Fischer, K., Betsch, T., & Haberstroh, S. (2005). Using expertise and experience for giving and taking advice. *The routines of decision making*, 157-173.
- Kapogiannis, D., Deshpande, G., Krueger, F., Thornburg, M. P., & Grafman, J. H. (2014). Brain networks shaping religious belief. *Brain Connect*, 4(1), 70-79.
- Krach, S., Hegel, F., Wrede, B., Sagerer, G., Binkofski, F., & Kircher, T. (2008). Can Machines Think? Interaction and Perspective Taking with Robots Investigated via fMRI. *PloS one*, 3(7), e2597.

- Kriegeskorte, N., Simmons, W. K., Bellgowan, P. S. F., & Baker, C. I. (2009). Circular analysis in systems neuroscience: the dangers of double dipping. *Nat Neurosci*, 12(5), 535-540.
- Krueger, F., Grafman, J., & McCabe, K. (2008). Neural correlates of economic game playing. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 363(1511), 3859-3874.
- Krueger, F., Landgraf, S., van der Meer, E., Deshpande, G., & Hu, X. (2011). Effective connectivity of the multiplication network: a functional MRI and multivariate Granger Causality Mapping study. *Hum Brain Mapp*, 32(9), 1419-1431.
- Kurth, F., Eickhoff, S. B., Schleicher, A., Hoemke, L., Zilles, K., & Amunts, K. (2010). Cytoarchitecture and probabilistic maps of the human posterior insular cortex. *Cereb Cortex*, 20(6), 1448-1461.
- Lacey, S., Stilla, R., Sreenivasan, K., Deshpande, G., & Sathian, K. (2014). Spatial imagery in haptic shape perception. *Neuropsychologia*, 60, 144-158.
- Lee, J., & Moray, N. (1992). Trust, control strategies and allocation of function in human-machine systems. *Ergonomics*, 35(10), 1243-1270.
- Madhavan, P., & Gonzalez, C. (2006). Effects of Sensitivity, Criterion Shifts, and Subjective Confidence on the Development of Automaticity in Airline Luggage Screening. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 50, 334-338.
- Madhavan, P., & Wiegmann, D. A. (2005). Cognitive anchoring on self-generated decisions reduces operator reliance on automated diagnostic aids. *Hum Factors*, 47(2), 332-341.
- Madhavan, P., & Wiegmann, D. A. (2007a). Effects of information source, pedigree, and reliability on operator interaction with decision support systems. *Hum Factors*, 49(5), 773-785.
- Madhavan, P., & Wiegmann, D. A. (2007b). Similarities and differences between human-human and human-automation trust: an integrative review. *Theoretical issues in ergonomics science*, 8(4), 277-301.
- Mars, R. B., Sallet, J., Schuffelgen, U., Jbabdi, S., Toni, I., & Rushworth, M. F. (2012). Connectivity-based subdivisions of the human right "temporoparietal junction area": evidence for different areas participating in different cortical networks. *Cereb Cortex*, 22(8), 1894-1903.
- Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An Integrative Model of Organizational Trust. *The Academy of Management Review*, 20(3), 709-734.
- McBride, S. E., Rogers, W. A., & Fisk, A. D. (2014). Understanding human management of automation errors. *Theoretical issues in ergonomics science*, 15(6), 545-577.
- Menon, V. (2011). Large-scale brain networks and psychopathology: a unifying triple network model. *Trends Cogn Sci*, 15(10), 483-506.
- Merritt, S. M., Heimbaugh, H., LaChapell, J., & Lee, D. (2013). I Trust It, But I Don't Know Why : Effects of Implicit Attitudes Toward Automation on Trust in an Automated System. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 55(3), 520-534.

- Meshi, D., Biele, G., Korn, C. W., & Heekeren, H. R. (2012). How expert advice influences decision making. *PloS one*, 7(11), e49748.
- Mitchell, J. P., Macrae, C. N., & Banaji, M. R. (2006). Dissociable medial prefrontal contributions to judgments of similar and dissimilar others. *Neuron*, 50(4), 655-663.
- Oldfield, R. C. (1971). The Assessment And Analysis Of Handedness: The Edinburgh Inventory. *Neuropsychologia*, 9, 97-113.
- Oliver, R. L. (1980). A Cognitive Model of the Antecedents and Consequences of Satisfaction Decisions. *Journal of Marketing Research*, 17(4), 460-469.
- Parasuraman, A. (2000). Technology Readiness Index (Tri): A Multiple-Item Scale to Measure Readiness to Embrace New Technologies. *Journal of Service Research*, 2(4), 307-320.
- Pearson, J. M., Heilbronner, S. R., Barack, D. L., Hayden, B. Y., & Platt, M. L. (2011). Posterior cingulate cortex: adapting behavior to a changing world. *Trends Cogn Sci*, 15(4), 143-151.
- Preusse, F., van der Meer, E., Deshpande, G., Krueger, F., & Wartenburger, I. (2011). Fluid intelligence allows flexible recruitment of the parieto-frontal network in analogical reasoning. *Front Hum Neurosci*, 5, 22.
- Ress, D., & Heeger, D. J. (2003). Neuronal correlates of perception in early visual cortex. *Nat Neurosci*, 6(4), 414-420.
- Rice, S., & McCarley, J. S. (2011). Effects of Response Bias and Judgment Framing on Operator Use of an Automated Aid in a Target Detection Task. *Journal of Experimental Psychology: Applied*, 17(4), 320-331.
- Riley, V. (1996). Operator reliance on automation: Theory and data.
- Roebroek, A., Formisano, E., & Goebel, R. (2005). Mapping directed influence over the brain using Granger causality and fMRI. *Neuroimage*, 25(1), 230-242.
- Rotter, J. B. (1967). A new scale for the measurement of interpersonal trust. *Journal of Personality*, 35(4), 651-665.
- Sathian, K., Deshpande, G., & Stilla, R. (2013). Neural changes with tactile learning reflect decision-level reweighting of perceptual readout. *The Journal of neuroscience*, 33(12), 5387-5398.
- Sathian, K., Lacey, S., Stilla, R., Gibson, G. O., Deshpande, G., Hu, X., . . . Glielmi, C. (2011). Dual pathways for haptic and visual perception of spatial and texture information. *Neuroimage*, 57(2), 462-475.
- Saxe, R., & Kanwisher, N. (2003). People thinking about thinking people: The role of the temporo-parietal junction in "theory of mind". *Neuroimage*, 19(4), 1835-1842.
- Singh, I. L., Molloy, R., & Parasuraman, R. (1997). Automation-induced "complacency": developement of the complacency-potential rating scale. *The International Journal of Aviation Psychology*, 3(2), 111-122.
- Skitka, L. J., Mosier, K., & Burdick, M. D. (2000). Accountability and automation bias. *International Journal of Human-Computer Studies*, 52(4), 701-717.
- Snizek, J. A., Schrah, G. E., & Dalal, R. S. (2004). Improving judgement with prepaid expert advice. *Journal of Behavioral Decision Making*, 17(3), 173-190.

- Sreenivasan, K. R., Havlicek, M., & Deshpande, G. (2015). Nonparametric hemodynamic deconvolution of fMRI using homomorphic filtering. *IEEE Trans Med Imaging*, 34(5), 1155-1163.
- Staudinger, M. R., & Buchel, C. (2013). How initial confirmatory experience potentiates the detrimental influence of bad advice. *Neuroimage*, 76, 125-133.
- Stilla, R., Deshpande, G., LaConte, S., Hu, X., & Sathian, K. (2007). Posteromedial parietal cortical activity and inputs predict tactile spatial acuity. *J Neurosci*, 27(41), 11091-11102.
- Strenziok, M., Krueger, F., Deshpande, G., Lenroot, R. K., van der Meer, E., & Grafman, J. (2010). Fronto-parietal regulation of media violence exposure in adolescents: a multi-method study. *Social Cognitive and Affective Neuroscience*.
- Suen, V. Y. M., Brown, M. R. G., Morck, R. K., & Silverstone, P. H. (2014). Regional Brain Changes Occurring during Disobedience to “Experts” in Financial Decision-Making. *PloS one*, 9(1), e87321.
- Terasawa, Y., Fukushima, H., & Umeda, S. (2013). How does interoceptive awareness interact with the subjective experience of emotion? An fMRI study. *Hum Brain Mapp*, 34(3), 598-612.
- Transportation Safety Administration. (2014). <http://www.tsa.gov/traveler-information/advanced-imaging-technology-ait>
- Van Overwalle, F. (2009). Social cognition and the brain: a meta-analysis. *Hum Brain Mapp*, 30(3), 829-858.
- Van Swol, L. M., & Sniezek, J. A. (2005). Factors affecting the acceptance of expert advice. *Br J Soc Psychol*, 44(Pt 3), 443-461.
- Wang, Y., & Quadflieg, S. (2015). In our own image? Emotional and neural processing differences when observing human-human vs human-robot interactions. *Soc Cogn Affect Neurosci*.
- Wen, X., Rangarajan, G., & Ding, M. (2013). Is Granger Causality a Viable Technique for Analyzing fMRI Data? *PloS one*, 8(7), e67428.
- Wheelock, M. D., Sreenivasan, K. R., Wood, K. H., Ver Hoef, L. W., Deshpande, G., & Knight, D. C. (2014). Threat-related learning relies on distinct dorsal prefrontal cortex network connectivity. *Neuroimage*, 102 Pt 2, 904-912.
- Xia, M., Wang, J., & He, Y. (2013). BrainNet Viewer: A Network Visualization Tool for Human Brain Connectomics. *PloS one*, 8(7), e68910.

## **CHAPTER THREE: THE IMPACT OF MISSES ON ADVICE UTILIZATION**

### **3.1 Abstract**

Our objective was to reveal the underlying neural mechanisms during advice utilization from expert human and machine agents with fMRI and multivariate Granger causality analysis. As society becomes more reliant on machines and automation, understanding how people utilize advice is a necessary endeavor. The impact of misses on decision-making and the neural basis involved with advice taking needs further exploration.

During the X-ray luggage-screening task, participants accepted or rejected good or bad advice from either the human or machine agent framed as experts with manipulated reliability (high miss rate). We showed that unreliable advice decreased performance and the machine-agent group decreased their advice utilization compared to the human-agent group. The differences in behaviors during advice utilization could be accounted for by high expectations of reliable advice and differences in attention allocation due to miss errors. Areas involved with the salience and mentalizing networks, as well as sensory processing involved with attention, were recruited during the task. The advice utilization network consisted of attentional modulation of sensory information with the lingual gyrus as the driver during the decision phase and the fusiform gyrus as the driver during the feedback phase. Our behavioral and fMRI results provide evidence demonstrating that miss errors from agents framed as experts decrease advice utilization due to reevaluation

of expectations. Assessment of the behavioral and neural mechanisms during unreliable advice can expand on the existing literature on miss errors, while also providing a neural network involved with advice utilization from humans and machines.

**This work is under review at *Human Factors*.**

### **Authors**

Kimberly Goodyear, Raja Parasuraman, Sergey Chernyak, Ewart de Visser, Poornima Madhavan, Gopikrishna Deshpande, Frank Krueger

### **Author Contributions**

K.G. and S.C. acquired the data for analysis. K.G., R.P. and F.K. contributed to the conception of the design. K.G., R.P., S.C., P.M., G.D. and F.K. contributed to interpretation of the data. K.G., R.P., S.C., E.D.V., P.M., G.D. and F.K. contributed to drafting of the work and revising it critically. K.G., R.P., S.C., E.D.V., P.M., G.D. and F.K. approved the final version to be published. K.G., R.P., S.C., E.D.V., P.M., G.D. and F.K. agreed to be accountable for all aspects of the work.

### **Funding**

This work was supported by the Air Force of Scientific Research [202857].

### **3.2 Introduction**

People are often given numerous options regarding the type and source of advice they can receive. For example, when individuals travel to a new country, they can ask a native citizen or use a smartphone with a Global Positioning System (GPS) for directions. Given the different options available, it is becoming a necessity to understand how individuals utilize or discount advice from different sources. Factors such as source credibility (expert and novice) (Madhavan & Wiegmann, 2007; Van Swol & Sniezek, 2005) and initial expectations of reliable advice (Dzindolet, Pierce, Beck, & Dawe, 2002) can influence how someone responds to advice. Dzindolet et al. (2002) proposed that individuals may possess a “perfect automation schema,” which is an expectation that automation performs near perfectly and can ultimately cause a person to disuse the advice given to them when errors occur. Initial expectations of reliable advice can be impacted, however, when disconfirmation evidence of misleading advice is encountered.

To fully understand the influence of bad advice on decision-making behaviors requires an examination of error types: false alarms and misses. The type of error is of particular interest because, while a false alarm error is misleading, it is not necessarily harmful. In contrast, a miss error can lead to disastrous results such as a luggage-screener failing to detect a bomb in a suitcase. Previous evidence has shown that false alarms can cause a “cry wolf effect,” in which an individual may tend to ignore true alerts (Breznitz, 2013) and misses may affect monitoring strategies leading to an adaptation of attention allocation (Onnasch, Ruff, & Manzey, 2014). False alarms have been shown to decrease trust and decrease reliance and compliance, while misses have been shown to

only decrease reliance (Dixon, Wickens, & McCarley, 2007; Rice & McCarley, 2011). Furthermore, studies comparing humans and machines have shown that expert humans were trusted more than expert machines due to differences in dispositional credibility (Madhavan & Wiegmann, 2007) and allocation of tasks to humans compared to automation can be affected by trust in automation (Lewandowsky, Mundy, & Tan, 2000). To expand on the existing literature on humans and machines, we previously investigated the impact of false alarms on decision-making behaviors (Goodyear et al., 2015, submitted), and to elaborate on those findings, the current study examined misses.

The neural processes involved with advice taking have been recently investigated with functional magnetic resonance imaging (fMRI) advice-taking paradigms examining expert advice (Boorman, O'Doherty, Adolphs, & Rangel, 2013; Meshi, Biele, Korn, & Heekeren, 2012) and during adaptive learning (Biele, Rieskamp, Krugel, & Heekeren, 2011). Furthermore, neuroimaging studies examining interactions between humans and robots during perspective taking (Krach et al., 2008) and during social observations (Wang & Quadflieg, 2015) have also been investigated. The default-mode network (e.g., temporoparietal junction, precuneus) and the salience network (dorsal anterior cingulate cortex, insulae) have been additionally implicated in other advice-taking tasks (Engelmann, Capra, Noussair, & Berns, 2009), as well as during robot-human interaction paradigms (Chaminade et al., 2012). However, in spite of the existing literature on advice taking, the neural basis and underlying brain networks associated with miss errors from expert human and machines remains to be elucidated.

We implemented an X-ray luggage-screening task with fMRI combined with multivariate Granger causality analysis (GCA) to investigate the impact of misses on decision-making behaviors and to reveal the underlying brain network associated with advice utilization from unreliable agents framed as experts. Based upon previous studies investigating misses and false alarms (Dzindolet et al., 2002; McBride, Rogers, & Fisk, 2014), we first hypothesized that unreliable advice would decrease performance (i.e., accuracy) compared to no advice. Furthermore, we expected advice utilization to decrease due to the significance of a miss error and due to disconfirmation evidence about the agents' expertise provided by feedback. We expected the reevaluation of the agents' perceived credibility to cause a mismatch of perceptions due to advice-incongruencies, which would ultimately cause an adjustment in attention allocation strategies. In addition, based upon previous work investigating advice acceptance and trust between expert human and machine agents (Madhavan & Wiegmann, 2007), we expected participants interacting with the machine agent to have a greater depreciation of advice utilization compared to the human agent due to perceptions involved with the perfect automation schema and varying degrees of perceived dispositional credibility. Brain regions involved with self-processing (e.g., precuneus) and error monitoring and salience detection (e.g., anterior cingulate cortex) would be recruited when comparing the human agent and the machine agent due to deviations in expectations (agents framed as experts), resulting from a change in attention strategies from a high miss rate.

### **3.3 Methods**

#### **Subjects**

A normative rating study and behavioral study were conducted at George Mason University (GMU) and an fMRI study was conducted at Auburn University (AU). All studies were conducted according to the ethical guidelines and principles of the Declaration of Helsinki. For the normative rating study, twenty-three male students (age ( $M \pm SD$ ) =  $24.0 \pm 2.6$ ) participated to standardize the X-ray luggage images for the experimental studies. For the behavioral study, twelve volunteers (7 males, 5 females; age =  $20.9 \pm 3.4$ ) participated to complete an X-ray luggage-screening task without receiving advice. For the fMRI study, twenty-four healthy right-handed volunteers (14 males, 10 females; age =  $22.3 \pm 2.4$ ) participated in the X-ray luggage-screening task while receiving advice. Participants gave written consent approved by the Institutional Review Boards at GMU and AU and they received financial compensation for their participation (see Goodyear et. al, 2015, submitted, for details on methods).

#### **X-ray Luggage-Screening Task**

Participants partook in an X-ray luggage-screening task and were asked to search for the presence or absence of a knife (Madhavan & Gonzalez, 2006) ([Appendix B.1a](#)). In the behavioral study, participants performed the task unassisted without receiving advice (no agent group). In the fMRI study, participants were assigned to either the human-agent group or the machine-agent group with 60% reliability and they received good (advice-congruent) and bad (advice-incongruent) advice ([Appendix B.1b](#)).

The jitter times were generated by an fMRI simulator software (<http://www.mccauslandcenter.sc.edu/CRNL/tools/fmrisim>) and consisted of a minimum of one second and an average of four seconds to optimize timing. Participants responded by using fiber optic response pads (Current Designs, <http://www.curdes.com/>); they were given an initial endowment of \$40 and each incorrect answer resulted in a deduction of \$0.30 from the remaining total. Performance, advice utilization, response times and monetary deductions were collected during the experiment. The stimuli were presented using E-Prime 2.0 (Psychology Software Tools, Inc.).

## **Procedure**

***Pre-Experimental Phase.*** Participants completed self-report questionnaires as control measures to investigate individual differences approximately one to two weeks before the fMRI experiment. The control measures included: Interpersonal Reactivity Index (IRI) (Davis, 1983), Complacency-Potential Rating Scale (CPS) (Singh, Molloy, & Parasuraman, 1997), National Readiness Technology Scale (NTRS) (Parasuraman, 2000), NEO Five-Factor Inventory (NEO-FFI) (Costa & McCrae, 1992), and Propensity to Trust (PTT) (Merritt, Heimbaugh, LaChapell, & Lee, 2013).

***Experimental Phase.*** Participants completed a practice run where they read descriptions about the human or machine agent (reliability was not disclosed), rated their trust in and reliability of the human or machine agent on a 10-point Likert scale (0 = very low, 10 = very high), familiarized themselves with the five possible knives that could be present in

the bags and then completed four practice trials of the task. The participants then completed two experimental runs of the task while in the scanner and rated reliability and trust afterwards.

***Post-Experimental Session.*** After completion of the fMRI experiment, participants were asked to rate their confidence in finding the target (i.e., knife) in each of the images presented during the experiment on a 10-point Likert scale (1 = very low, 10 = very high).

### **Neuroimaging Acquisition**

Imaging data were acquired on a 7T actively shielded whole-body scanner (Siemens Magnetom) with a 32-channel head coil (Nova Medical) at AU MRI Research Center, Auburn, Alabama. The anatomical imaging data were based on a 3D T1-weighted MPRAGE sequence with TR = 2020 ms, TE = 2.7 ms, flip angle = 7°, slice thickness = 1.2 mm, voxel dimension = 1.1 mm x 1.1 mm x 1.2 mm and number of slices = 240. The functional imaging data were based on a 2D gradient-echo multiband EPI sequence with TR = 1000 ms, TE = 20 ms, flip angle = 70°, slice thickness = 2 mm, voxel dimensions = 2.1 mm x 2.1 mm x 2.0 mm, number of slices = 45 per volume in an axial orientation parallel to the anterior-posterior commissure and a multiband factor of 2. The first two volumes were discarded to allow for T1 equilibrium effects and a total of 660 volumes were taken for each run.

### **Behavioral Data Analysis**

Behavioral data was analyzed with the Statistical Package for the Social Sciences 20.0 (SPSS 20.0, IBM Corp.) and the alpha was set to  $p < .05$  (two-tailed). Data were normally distributed (Kolmogorov–Smirnov test) and assumptions for analyses of variance (Bartlett’s test) were not violated. To investigate task performance between the agents and the no agent group, a one-way analysis of variance (ANOVA) with Agent (human, machine, no agent) as the between-subjects factor. Mixed 2 x 2 x 2 repeated-measures ANOVAs with Advice (good, bad) and Time (run 1, run 2) as within-subjects factors and Agent (human, machine) as the between-subjects factor were employed to examine advice utilization, response times and monetary deductions. In addition, we investigated reliability, trust and confidence ratings with mixed 2 x 2 repeated-measures ANOVAs with Agent (human, machine) as the between-subjects factor. The within-subjects factor for the reliability/trust ratings were Time (pre, post) and for confidence ratings was Target (yes, no).

### **Neuroimaging Data Analysis**

The fMRI data was analyzed through NeuroElf software (<http://neuroelf.net>) and BrainVoyager QX 2.8 (Brain Innovation). The functional imaging data were preprocessed using Statistical Parametric Mapping (SPM, Wellcome Department of Cognitive Neurology) functions batched via NeuroElf, including three-dimensional motion correction (six parameters), slice-scan time correction (temporal interpolation). A mean functional image was computed for each participant across all runs and was then

co-registered with the anatomical images using a joint-histogram for the different contrast types. Preprocessing procedures for the anatomical images included segmenting images with a unified segmentation procedure (Ashburner & Friston, 2005) and the functional images had spatial warping applied to them to normalize the data to a standard Montreal Neurological Institute (MNI) brain template. To account for any residual differences across participants, spatial smoothing (Gaussian filter of 6 mm FWHM) was applied to the images.

A general linear model (GLM) that was corrected for first-order serial correlations fit to the data (Friston, Harrison, & Penny, 2003), which consisted of thirty-seven regressors based on advice utilization (accept, reject), advice type (good, bad), time (run 1, run 2) for each of the five phases (fixation, advice, bag, decision, feedback) and seven parametric regressors of no interest for the global signal and 3D motion correction (translations in X, Y, Z directions, rotations around X, Y, Z axes). The regressor time courses were adjusted for the hemodynamic response delay by convolution with a dual-gamma canonical hemodynamic response function (Buckel, Holmes, Rees, & Friston, 1998). Random-effect analyses were performed at the multi-subject level to explore brain activations associated with the decision and feedback phases during advice utilization.

Mixed 2 x 2 x 2 ANOVAs on parameter estimates were applied with Advice (good, bad) and Time (run 1, run 2) as within-subjects factors and Agent (human, machine) as the between-subjects factor. Brain activations for the decision and feedback phases were reported after correcting for multiple comparisons using a cluster-level

statistical threshold (Cluster-level Statistical Threshold Estimator plugin in BrainVoyager QX). The thresholded map ( $p < .005$ ) was used for a whole-brain correction criterion, which is based off an estimate of the map's spatial smoothness and on a Monte Carlo simulation (1,000 iterations). The minimum cluster size at a specified confidence level ( $\alpha = 0.05$ ) was then calculated (Forman et al., 1995; Goebel, Esposito, & Formisano, 2006). The significant activation clusters were displayed in MNI space on an anatomical brain template reversed left to right (i.e., radiological convention).

### **Effective Connectivity Analysis**

Effective (or directional) connectivity data were analyzed using a code developed in-house using MATLAB ([www.mathworks.com](http://www.mathworks.com)) (Grant et al., 2014; Lacey, Stilla, Sreenivasan, Deshpande, & Sathian, 2014) (for more details on methods see Appendix B.2). The effective connectivity in the network of activated regions was performed through multivariate Granger causality analysis (GCA) and only regions that survived the fMRI analysis threshold for the main effect of Agent (human, machine) for the decision and feedback phases were selected as ROIs. Time series of the blood-oxygen-level-dependent (BOLD) signal for the selected ROIs were extracted around peak activation maxima (sphere of  $6 \times 6 \times 6 \text{ mm}^3$ ), averaged across voxels and normalized across participants, per run. Blind hemodynamic deconvolution of the mean ROI BOLD time series was performed using a Cubature Kalman filter and smoother (Havlicek, Friston, Jan, Brazdil, & Calhoun, 2011) and the resulting latent neural signals were entered into a first order dynamic multivariate autoregressive (dMVAR) model to assess directed

interactions of multiple nodes as a function of time (Feng et al., 2015; Grant, Wood, Sreenivasan, Wheelock, & White, 2015; Hampstead, Khoshnoodi, Yan, Deshpande, & Sathian, 2016; Hutcheson et al., 2015; Wheelock et al., 2014).

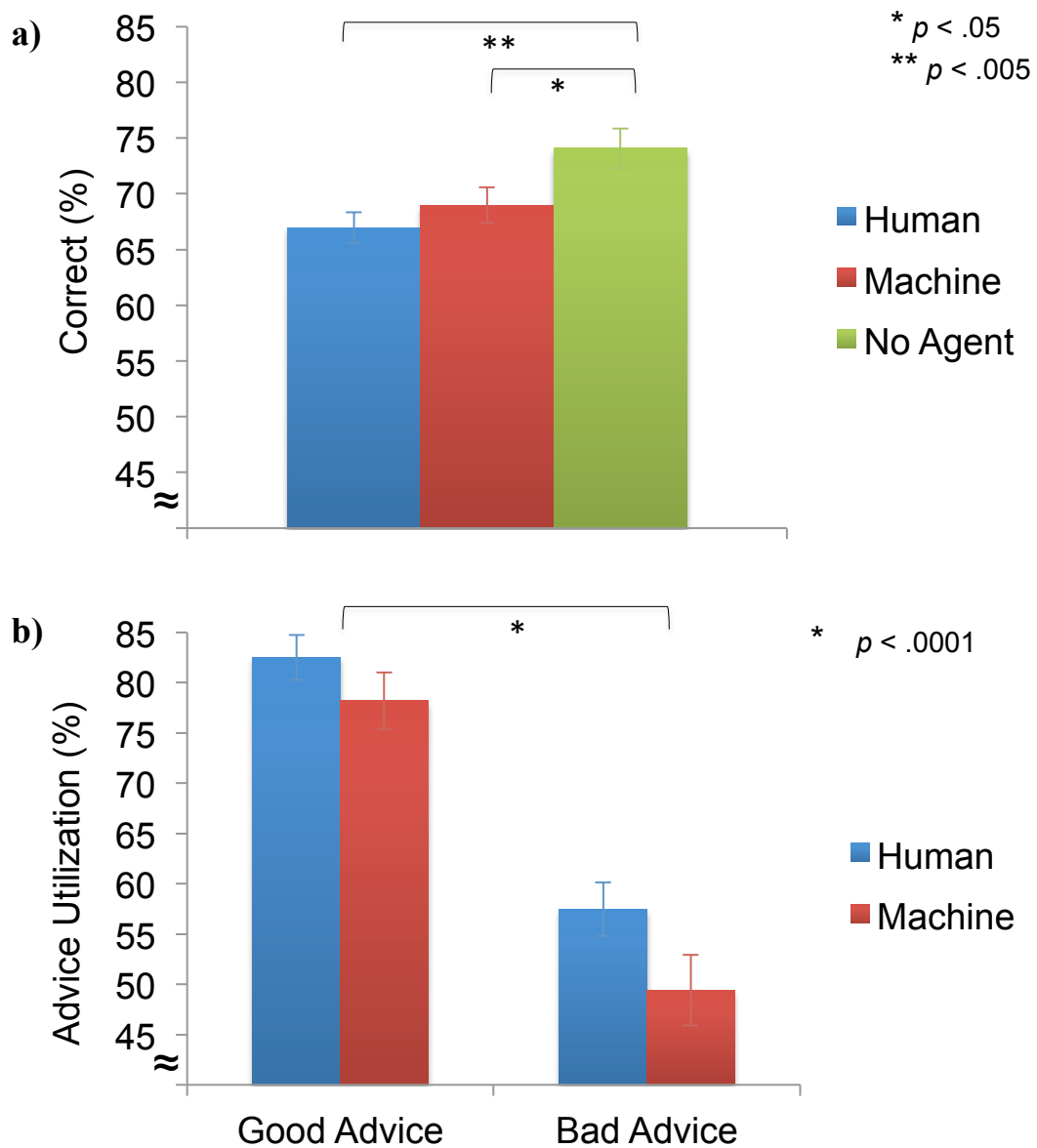
Granger connectivity path weights for the condition of interest (advice utilization) for each agent (human, machine) were extracted, populated into two samples, and independent samples *t*-tests were employed ( $q(\text{FDR}) < .05$ ) (Benjamini & Hochberg, 1995) to reveal significantly different effective connectivity paths between the agent groups ([Appendix B.3](#)). Effective connectivity of brain regions (i.e., nodes, edges) was displayed on a brain surface using BrainNet Viewer, a graphical interface visualization tool (Xia, Wang, & He, 2013).

### **3.4 Results**

#### **Behavioral Results**

The one-way ANOVA comparing performance between the agent groups and the no agent group revealed a significant main effect of Agent ( $F(2, 33) = 5.77, p = .007$ ).

Planned follow-up analysis revealed that the no agent group performed better than the human-agent group ( $t(22) = -3.37, p = .003$ ) and the machine-agent group ( $t(22) = -2.24, p = .035$ ) ([Fig. 7a](#)).



**Figure 7. Miss Behavioral Results**

**Results for the Decision Phase ( $M \pm SEM$ ).** **a) Task Performance.** The no agent group performed better than human- and machine-agent groups. **b) Advice Utilization.** Advice utilization was significantly lower for bad advice compared to good advice and was also significantly lower for the machine-agent group compared to the human-agent group.

Next, we looked at advice utilization by implementing mixed ANOVAs. For *advice utilization*, significant main effects of Agent ( $F(1, 22) = 5.24, p = .032$ ), Advice ( $F(1,22) = 140.72, p < .0001$ ) and Time ( $F(1,22) = 22.36, p < .0001$ ) were found. These results indicate that participants accepted advice more from the human agent compared to the machine agent. Furthermore, good advice was accepted more than bad advice and advice utilization decreased over time (Fig. 7b). In addition, a significant two-way interaction of Advice x Time was identified ( $F(1, 22) = 10.17, p = .004$ ), but no significant two-way interaction effects of Advice x Agent ( $F(1, 22) = 0.69, p = .415$ ), Time x Agent ( $F(1, 22) = 0.46, p = .505$ ), or three-way interaction of Advice x Time x Agent ( $F(1, 22) = 1.40, p = .249$ ) were found.

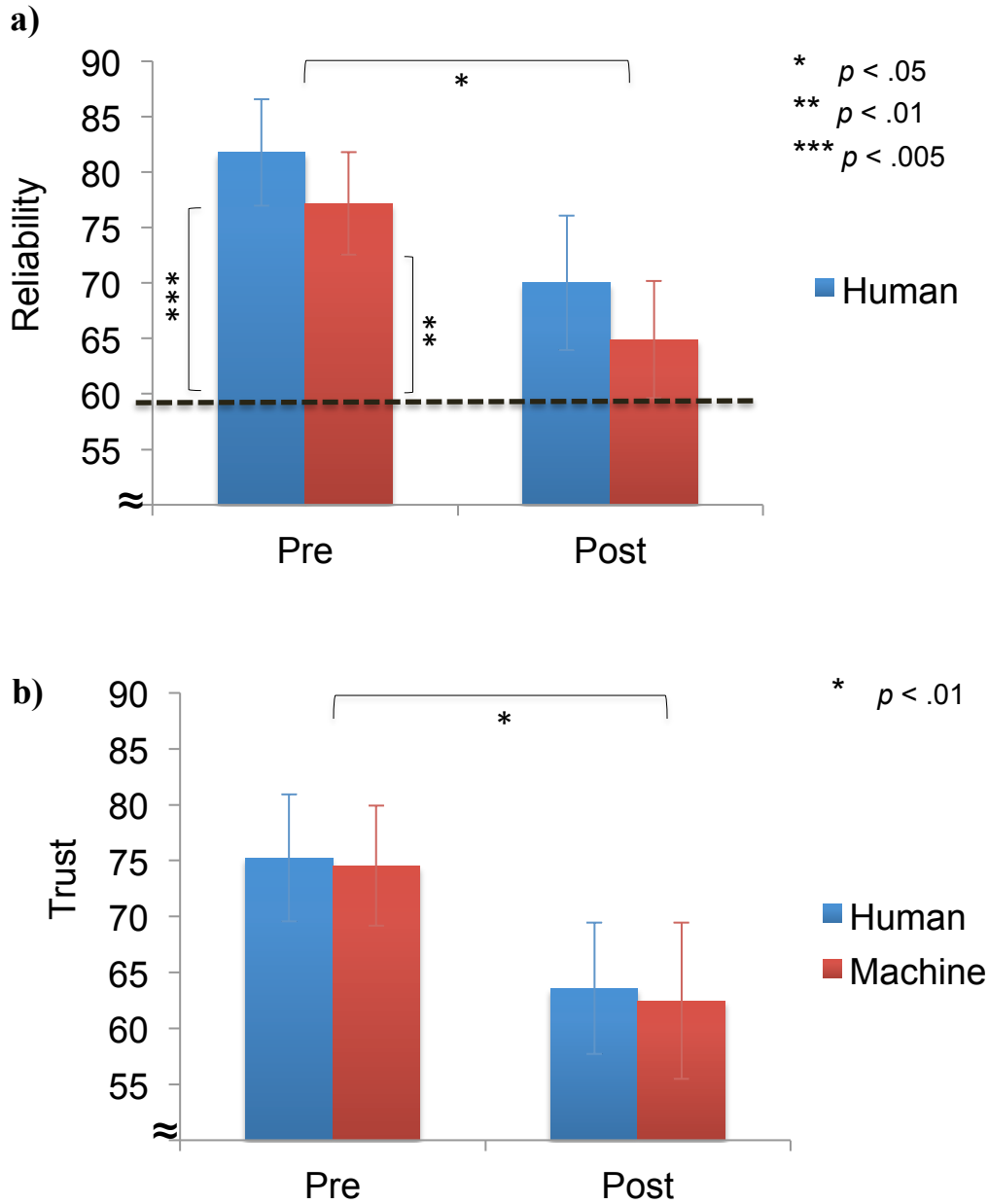
In addition, we looked at pre- and post-reliability/trust ratings. One participant's data were not used due to lack of understanding, which was indicated by the high values for all pre/post scales. The *reliability ratings* showed no significant main effect of Agent ( $F(1, 21) = 0.76, p = .394$ ), but a significant main effect of Time ( $F(1, 21) = 5.43, p = .030$ ), showing that reliability ratings decreased from pre- to post-experiment (Fig. 8a). No significant interaction effect of Time x Agent ( $F(1, 21) = 0.00, p = .960$ ) was found. Furthermore, one-sample *t*-tests on perceived versus actual reliability (60%) of the agent revealed that pre-reliability ratings were significantly higher than the actual reliability for the human agent ( $t(11) = 4.53, p = .001$ ) and the machine agent ( $t(10) = 3.55, p = .005$ ). For *trust ratings*, no significant main effect of Agent ( $F(1, 21) = 0.01, p = .905$ ) was found, but a significant main effect of Time ( $F(1, 21) = 8.18, p = .009$ ) was observed, showing that trust ratings significantly decreased from pre- to post-experiment (Fig. 8b).

No significant interaction effect of Time x Agent ( $F(1, 21) = 0.00, p = .960$ ) was demonstrated.

We next analyzed differences in control measures (e.g., demographic measures and questionnaires) with independent samples *t*-tests. No significant group differences were identified for any of the control measures ([Appendix B.4](#)).

For *response times*, a significant main effect of Time ( $F(1, 22) = 5.42, p = .030$ ) was found, indicating that responses were faster during run 2 compared to run 1 ([Appendix B.5a](#)). No significant main effects of Agent ( $F(1, 22) = 0.77, p = .389$ ) or Advice ( $F(1, 22) = 1.34, p = .260$ ) were revealed and no significant interaction effects of Advice x Agent ( $F(1, 22) = 3.27, p = .084$ ), Time x Agent ( $F(1, 22) = 3.28, p = .084$ ), Advice x Time ( $F(1, 22) = 2.46, p = .131$ ) or Advice x Time x Agent ( $F(1, 22) = 0.73, p = .401$ ) were found.

For *monetary deductions*, a significant main effect of Time ( $F(1, 22) = 7.13, p = .014$ ) was revealed, indicating that deductions were higher during run 1 compared to run 2 ([Appendix B.5b](#)). No significant main effects of Advice ( $F(1, 22) = 1.34, p = .260$ ) and Agent ( $F(1, 22) = 0.69, p = .414$ ), or interaction effects of Advice x Agent ( $F(1, 22) = 3.54, p = .073$ ), Advice x Time ( $F(1, 22) = 0.08, p = .776$ ), Time x Agent ( $F(1, 22) = 0.66, p = .427$ ), or Advice x Time x Agent ( $F(1, 22) = 2.50, p = .128$ ) were found.



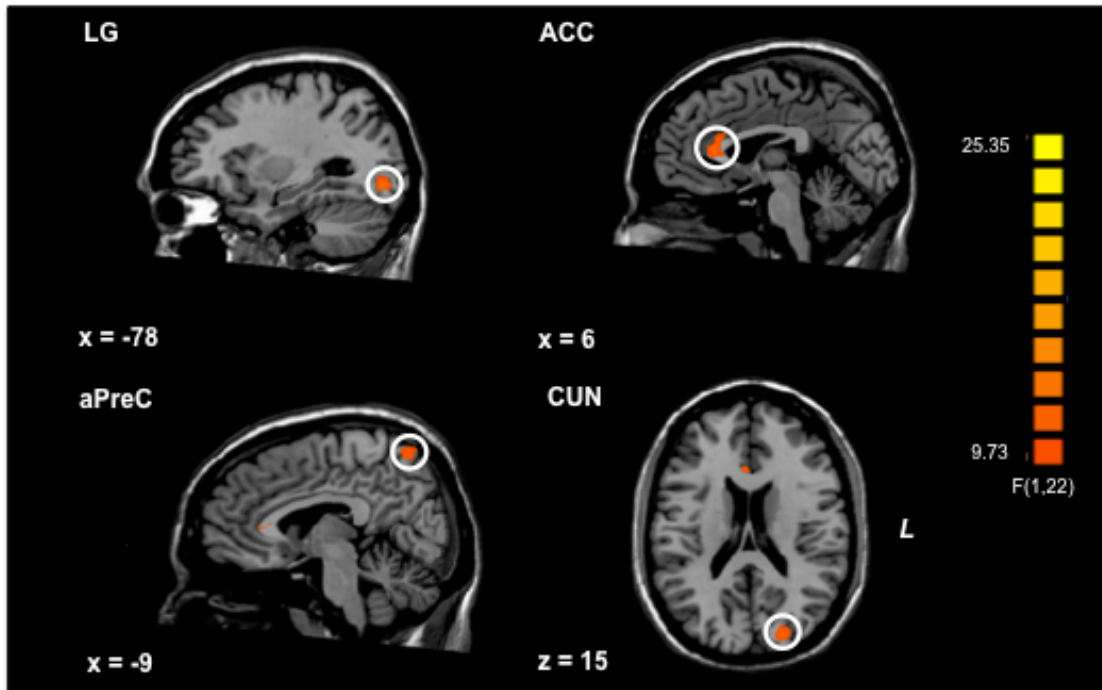
**Figure 8. Miss Rating Results**

**Results for Ratings ( $M \pm SEM$ ).** **a) Pre- and Post-Reliability.** For both groups, the perceived pre-reliability was significantly higher than the actually reliability of the agent (60%) and post-reliability ratings significantly decreased. **b) Pre- and Post-Trust.** Post-trust was significantly lower than pre-trust for both groups.

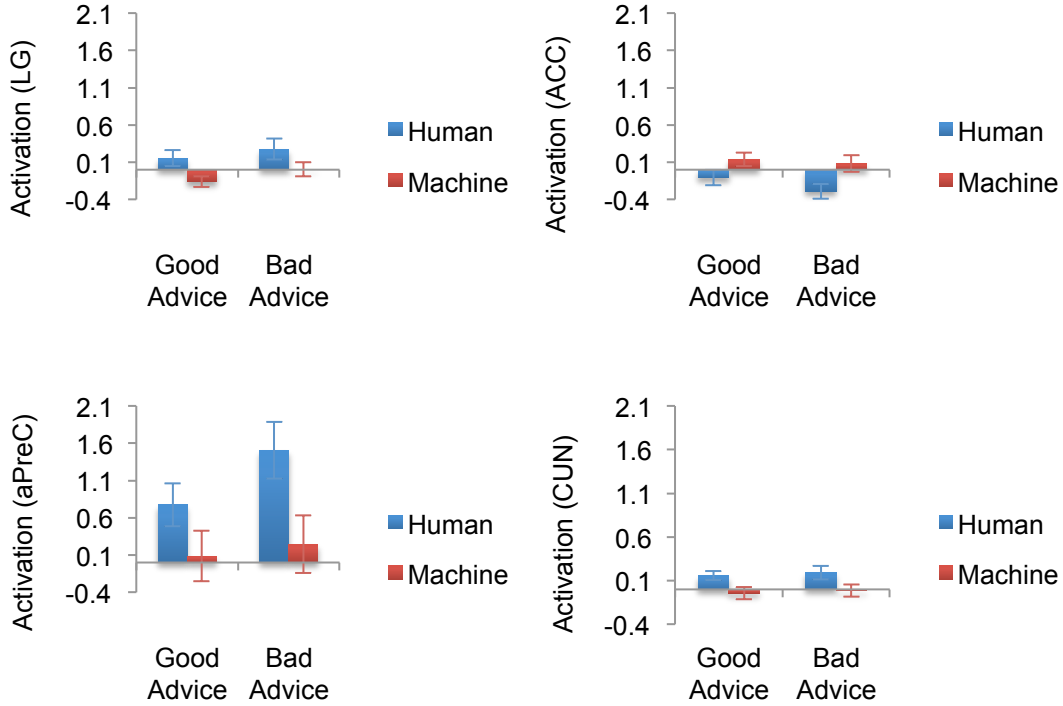
For *confidence ratings*, no main effect of Agent ( $F(1, 22) = 0.39, p = .538$ ) or significant interaction effect of Target x Agent ( $F(1, 22) = 0.50, p = .488$ ) was found, but a significant main effect of Target ( $F(1, 22) = 46.30, p < .0001$ ) was revealed, indicating that confidence was rated higher on target bags compared to non-target bags ([Appendix B.6](#)).

### Neuroimaging Results

We investigated brain activations during the decision and feedback phases with mixed ANOVAs. For the decision phase, a significant main effect of Agent ( $\alpha < .05, k = 11$ ) was found in the right (R) lingual gyrus (LG) (BA 18), R anterior cingulate cortex (ACC) (BA 24), left (L) anterior precuneus (aPreC) (superior parietal lobule; BA 7), and L cuneus (CUN) (BA 18) ([Fig. 9](#), [Fig. 10](#), [Tab. 3](#)). A main effect of Advice ( $\alpha < .05, k = 11$ ) was found in the R middle frontal gyrus (BA 8), R medial frontal gyrus (BA 8), R rostromedial prefrontal cortex (rPFC) (superior frontal gyrus; BA 10), R primary visual cortex (V1) (BA 17), R pre-supplementary motor area (pre-SMA) (superior frontal gyrus; BA 6), L cerebellar culmen, L inferior occipital gyrus (IOG) (BA 18).



**Figure 9. Miss Brain Activations During Decision Phase**  
( $\alpha < .05$ ,  $k = 11$ ). The main effect of Agent during the decision phase significantly activated the right lingual gyrus (LG), right anterior cingulate cortex (ACC), left anterior precuneus (aPreC) and left cuneus (CUN).



**Figure 10. Miss Activation Patterns During Decision Phase**

The activation pattern indicates higher activation for the human- compared to machine-agent group for all regions except the ACC. The bar plots shown are for visualization purposes. To avoid circularity, or double dipping, no further statistical analyses were performed for the decision and feedback phases (Kriegeskorte, Simmons, Bellgowan, & Baker, 2009).

**Table 3. Miss Brain Regions**

**Brain Regions Associated with the Agent and Advice Main Effects.** Brain regions showing significant activation clusters associated during the decision phase: Agent (minimum cluster of 11) and Advice (minimum cluster of 11); and feedback phase: Agent (minimum cluster of 10) and Advice (minimum cluster of 9) ( $\alpha < .05$ , cluster-level threshold corrected, MNI space).

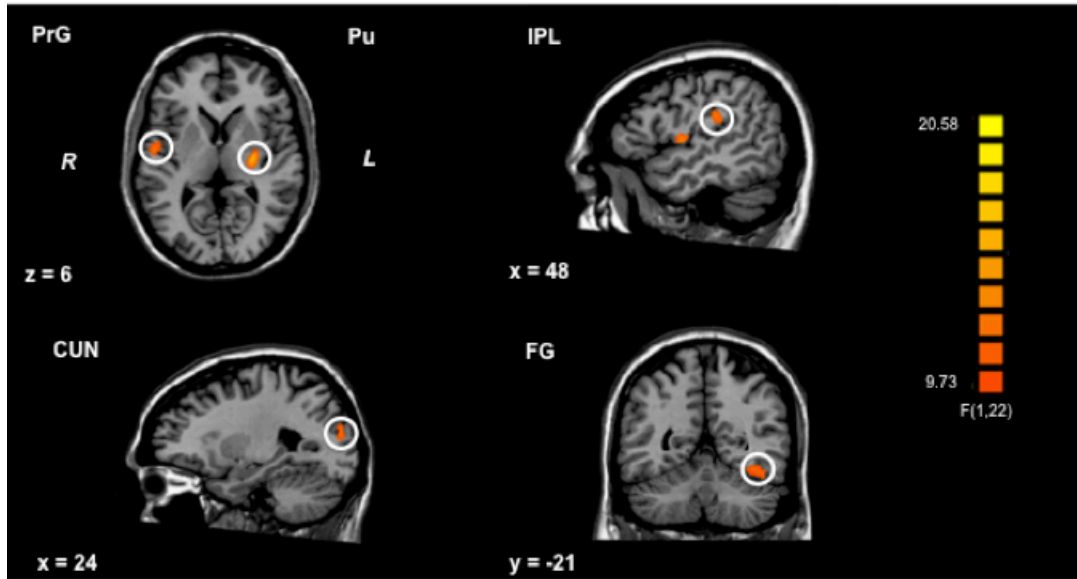
	<i>F</i> value	Cluster Size (mm <sup>3</sup> )	x	y	z
<b>Decision Phase</b>					
<i>Agent</i>					
Right lingual gyrus	18.44	629	27	-78	-6
Right anterior cingulate cortex	24.35	1246	6	27	12

Left anterior precuneus	19.84	727	-9	-63	57
Left cuneus	19.95	758	-21	-84	15
<b><i>Advice</i></b>					
Right middle frontal gyrus	24.75	822	42	18	42
Right medial frontal gyrus	21.3	3182	21	27	33
Right rostrolateral prefrontal cortex	28.51	560	24	54	6
Right primary visual cortex	19.72	1722	15	-96	-3
Right pre-supplementary motor area	19.86	665	6	9	56
Left cerebellar culmen	17.93	601	-12	-36	-24
Left inferior occipital gyrus	16.37	1936	-24	-90	-6
<b>Feedback Phase</b>					
<b><i>Agent</i></b>					
Right precentral gyrus	16.66	456	51	-6	6
Right inferior parietal lobule	15	398	48	-26	24
Right cuneus	15.37	422	24	-84	15
Left putamen	19.3	1445	-27	-15	6
Left fusiform gyrus	19.58	990	-42	-47	-21
<b><i>Advice</i></b>					
Right postcentral gyrus	19.33	960	42	-18	27
Right middle frontal gyrus	16.78	631	33	21	39
Right hippocampus	18.19	1347	29	-39	3
Right extra-nuclear	18.66	468	24	21	15
Right orbitofrontal cortex	25.94	892	21	45	-3
Right posterior cingulate cortex	31.47	1049	12	-63	23
Right anterior precuneus	23.25	1865	6	-69	47
Left cerebellar culmen	23.43	2945	-6	-42	-21
Left pons	16.91	373	3	21	51
Left pre-supplementary motor area	18.27	644	-18	-24	-30
Left parahippocampal gyrus	29.02	1102	-24	-42	0
Left postcentral gyrus	31.04	1300	-42	-21	27

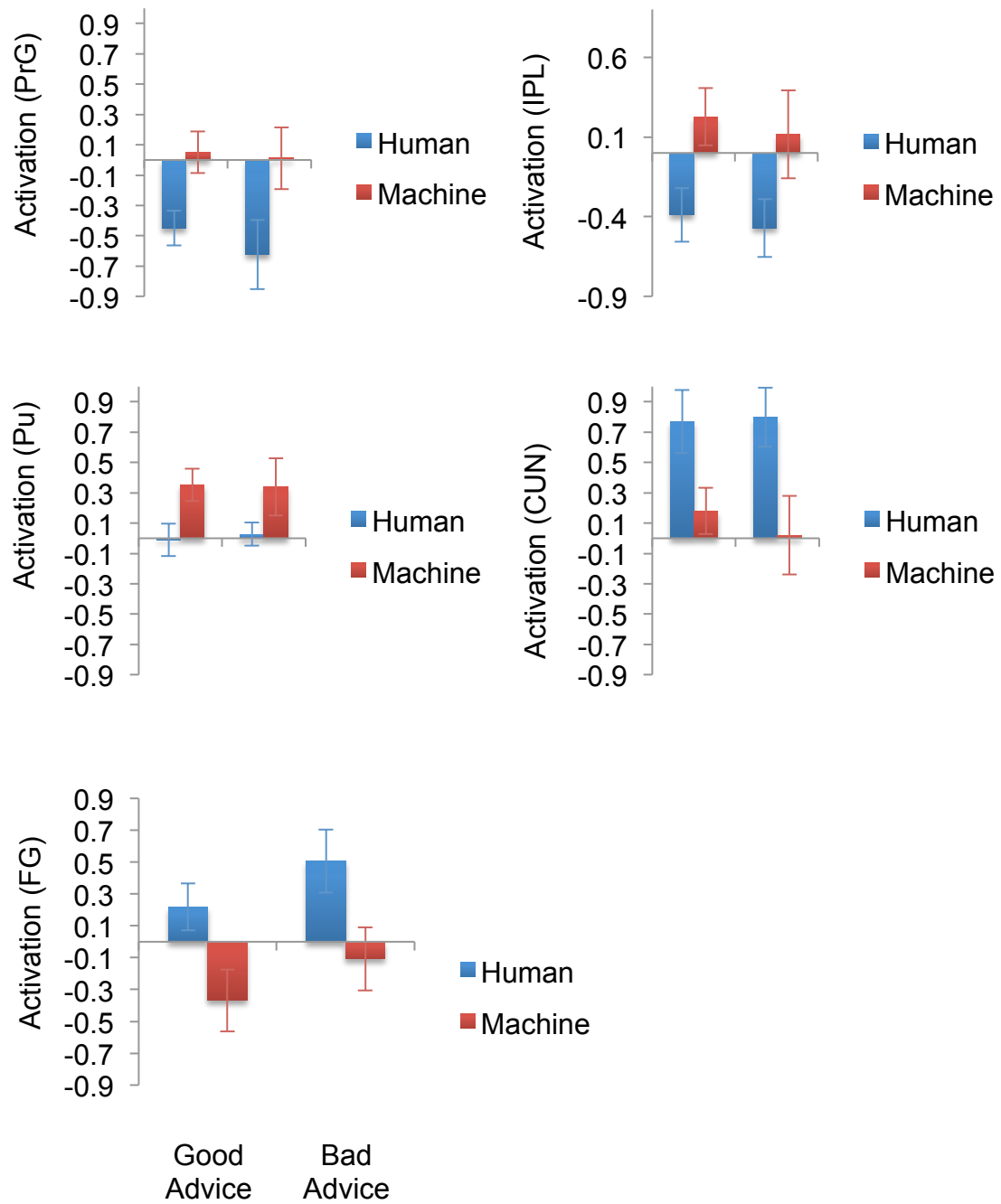
For the feedback phase, a main effect of Agent ( $\alpha < .05$ ,  $k = 10$ ) was found in the R precentral gyrus (PrG) (BA 6), R inferior parietal lobule (IPL) (BA 40), R CUN (BA 17), L putamen (Pu) and L fusiform gyrus (FG) (BA 37) ([Fig. 11](#), [Fig. 12](#), [Tab. 3](#)).

Lastly, a significant main effect of Advice ( $\alpha < .05$ ,  $k = 9$ ) during the feedback phase was

found in the R postcentral gyrus (PoG) (BA 3), R middle frontal gyrus (BA 8), R hippocampus, R extra-nuclear, R orbitofrontal cortex (OFC) (BA 10/11), R posterior cingulate cortex (PCC) (BA 31), R aPreC (BA 7), L cerebellar culmen, L pre-SMA (BA 6/8), L pons, L parahippocampal gyrus (BA 19) and L PoG (BA 2).



**Figure 11. Miss Brain Activations During Feedback Phase** ( $\alpha < .05$ ,  $k = 10$ ). The main effect of Agent during the feedback phase significantly activated the right precentral gyrus (PrG), right inferior parietal lobule (IPL), R cuneus (CUN), left putamen (Pu) and left fusiform gyrus (FG).

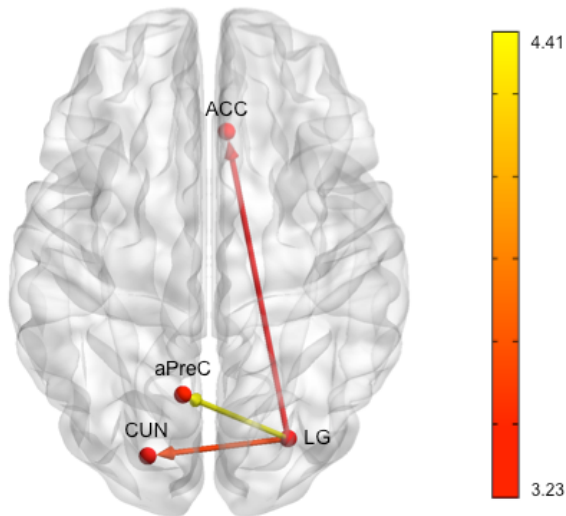


**Figure 12. Miss Brain Activation Patterns During Feedback Phase**

The activation pattern shows higher activation for the machine-agent group compared to the human-agent group for all regions except for FG and CUN. The bar plots shown are for visualization purposes.

## Effective Connectivity Results

To identify effective connectivity among brain regions when comparing the human to the machine agents during the decision and feedback phases, we implemented multivariate GCA based upon our results from the fMRI analysis ( $q(\text{FDR}) < .05$ ). The LG was identified as the source ROI for the advice utilization network for the decision phase, that sent output connections to all target ROIs (ACC, aPreC, CUN) and the FG was the source ROI for the feedback phase sending an output connection to the IPL (Fig. 13, Tab. 4).



**Figure 13. Miss Results for Multivariate Granger Causality Analysis**

The effective connectivity network for advice utilization during the decision phase when comparing the human with machine agent showed that the LG (lingual gyrus) was the driver of the network and source ROI, sending outputs to all target ROIs (ACC (anterior cingulate cortex), aPreC (anterior precuneus), and CUN (cuneus)) (all connections survived  $q(\text{FDR}) < .05$ ). The color bar represents the  $t$ -value of the comparisons shown in Table 4.

**Table 4. Miss Granger Causality Analysis**

**Path Weights for Granger Causality Analysis.** The path weights displayed show significant effective connectivity paths that are stronger in the human-agent group compared to the machine-agent group during advice utilization ( $q(\text{FDR}) < .05$ ). The directionality of the connectivity is shown in the first two columns, with the source column showing the ROIs that predict activation in the target column ROIs. The strength of connectivity is given by the mean path weights in the third column. LG, lingual gyrus; ACC, anterior cingulate cortex; aPreC, anterior precuneus; CUN, cuneus; FG, fusiform gyrus; IPL, inferior parietal lobule.

Source	Target	Path weight		<i>t</i> value	<i>p</i> value
		Human	Machine		
Decision Phase					
LG	ACC	0.087	-0.003	3.23	6.18 x 10 <sup>-6</sup>
	aPreC	0.115	0.009	4.41	5.23 x 10 <sup>-8</sup>
	CUN	0.094	-0.006	3.49	2.43 x 10 <sup>-8</sup>
Feedback Phase					
FG	IPL	0.087	-0.156	3.03	1.20 x 10 <sup>-4</sup>

### 3.5 Discussion

The objective of this research was to expand on our earlier work investigating the behavioral and neural signatures of advice utilization differences between expert human and machine agents during good and bad advice (Goodyear et al., 2015, submitted). We manipulated agent reliability with a high miss rate to reveal the underlying neural basis (in terms of both activated brain regions and the directional interactions between them) involved with advice utilization. We revealed that unreliable advice decreased performance overall as shown by other behavioral studies investigating human-machine interactions (Dzindolet et al., 2002; Goodyear et al., 2015, submitted), and advice utilization decreased more for the machine-agent group compared to the human-agent

group, coinciding with another study investigating the effects of source credibility with varying reliability from humans and machines (Madhavan & Wiegmann, 2007).

As hypothesized, our results demonstrated that advice utilization decreased more for the machine-agent group compared to the human-agent group. The degradation of advice utilization occurred regardless of the type of the advice (good, bad) given, showing that disconfirmation experience during bad advice had an effect on all decision-making behaviors. In our earlier work, we showed that false alarms caused a degradation of advice utilization during bad advice (Goodyear et al., 2015, submitted), but for our current study, we expected that misses would cause an overall adjustment in attention allocation due to previous evidence showing that more critical types of events (misses) lead to an adaptation in monitoring strategies (Onnasch et al., 2014). Our results indicated that advice utilization decreased for both groups, which provides evidence that participants made changes in their decision-making behaviors to compensate for the unreliable advice that they received.

In addition, we compared the pre-reliability ratings with the actual reliability of each agent to uncover any preconceived notions that participants had about the human and machine agents. We demonstrated that for both groups the pre-reliability ratings were significantly higher than the actual reliability, which could indicate that participants had high initial expectations of reliable advice since the agents were framed as experts. In addition, reliability ratings decreased overall from pre- to post-experiment, showing that participants were able to decipher the performance of the agents, while also recalibrating their expectations due to bad advice. Furthermore, we revealed that trust

decreased overall from pre- to post-experiment, revealing that misses degraded trust, which has previously been reported for false alarms (Dixon et al., 2007; Goodyear et al., 2015, submitted; Rice & McCarley, 2011). Although the reliability and trust ratings did not significantly decrease more for the machine-agent group, the ratings were still lower compared to the human-agent group, which could show that as trust and reliability decreased, advice utilization degraded as well. Lastly, since we showed no differences for control measures or confidence ratings between the agent groups, our results cannot be explained by those findings.

We next identified the neural mechanisms and the underlying directional brain network differentially involved with advice utilization between humans and machines. For the decision phase, our effective connectivity network revealed the LG as the driver, or source ROI, of the network, sending outputs to the ACC, aPreC and CUN. Furthermore, the strength of the paths emanating from LG were significantly higher for human advice compared to machine advice. The results indicate that the LG perceivably modulated attention during advice utilization through the bottom-up sensory processing of task-relevant information. It has been postulated that sensory processing involves a large-scale integration of networks with attention modulation to form a behavioral outcome, or a cognition (Mesulam, 1998). For example, it has been shown that detection of stimulus information initially starts in primary sensory areas, and is then conveyed to regions such as the ACC, showing the interaction between bottom-up and top-down processing during attentional control (Crottaz-Herbette & Menon, 2006). Furthermore, a study investigating advisor competence showed increased activity in the visual cortex

during advice integration from incompetent advisors (Schilbach, Eickhoff, Schultze, Mojzisch, & Vogeley, 2013). The authors conclude that the activity in the visual cortices may relate to “perceptually based strategies” during reassessment of one’s own judgments, which could support our findings about the influence of visual regions on upstream structures such as PreC and ACC during advice utilization with unreliable human advisors. Moreover, the involvement of the visual areas during the decision phase could be attributed to the fact that participants had to revisualize the X-ray images in order to compare what they saw to the advice they received.

Furthermore, our neuroimaging results for the decision phase revealed brain regions associated with attentional control and salience detection (ACC), self-processing (aPreC) and sensory information processing (LG and CUN). LG activation has been associated with comparing advice versus no advice in expert and peer groups (Suen, Brown, Morck, & Silverstone, 2014) and activity in the LG and CUN has been implicated during decisions under risk when comparing a message to accept or reject advice with no message (Engelmann et al., 2009) and during decisions correlated with value or saliency (Litt, Plassmann, Shiv, & Rangel, 2011). ACC activation has been shown to be involved with conflict monitoring during decision-making (Botvinick, 2007) and error detection and prediction error (Beckmann, Johansen-Berg, & Rushworth, 2009), while the PreC has been identified to play a role in integrations of one’s mental state (Terasawa, Fukushima, & Umeda, 2013). Our neuroimaging results demonstrated that all areas except for the ACC had higher activations for the human-agent group compared to the machine-agent group, indicating that participants in the human-agent

group may have had a greater increase in perceptual processing and perceivably less monitoring of errors. Conversely, participants in the machine-agent group were more attuned to the advice errors, which was also indicated behaviorally, which could explain the ACC activation differences.

In addition to the decision phase, we expected a behavioral adjustment in advice utilization due to feedback. For the feedback phase, our effective connectivity network showed that the FG was the driver of the network that sent an output to the IPL. The FG has been associated with receipt of monetary rewards and penalties during an outcome phase (Dillon et al., 2008), while the IPL has been identified to play a role during advice evaluation when interacting with competent and incompetent advisors (Schilbach et al., 2013) and during decision uncertainty when given trial-by-trial feedback (Vickery & Jiang, 2009). Furthermore, the neuroimaging results for the feedback phase revealed activity in the PrG, CUN and Pu. Activity in the PrG has been implicated during comparisons of humans and computers during rock-paper-scissors games (Chaminade et al., 2012) and CUN activity has been shown to be related to inferential errors during a feedback phase (Cooper, Kreps, Wiebe, Pirkel, & Knutson, 2010). Lastly, we revealed activity in the dorsal striatum (Pu), which has been implicated in stimulus-response learning (Packard & Knowlton, 2002) and during responses to affective feedback in regards to valence and magnitude (Delgado, Locke, Stenger, & Fiez, 2003). Our results for the feedback phase illustrate that, for all regions except for CUN and FG, activations were higher for the machine-agent group compared to the human-agent group. This pattern of activation indicates that as participants in the machine-agent group became

more aware of the errors in advice, they may have placed more value on the outcome of their decisions as opposed to just processing of sensory information.

There are a couple limitations that need to be considered with the interpretation of our results. First, we looked at differences between good and bad advice with only misses as the type of error. However, our previous research on false alarms (Goodyear et al., 2015, submitted) provided substantiation for expanding on the effects of advice utilization with different error types and future studies could include both types of errors to compare the two directly. In addition, participants received advice before they made their decisions in order to prevent cognitive anchoring, or the tendency to rely on the first piece of information acquired. Future studies could investigate the effects of cognitive anchoring by implementing a task where participants receive advice after they make their decisions.

In conclusion, our results have shown that advice utilization differs between humans and machines and those distinctions are contingent on miss errors. Our findings expand on the existing literature by showing that misses degrade advice utilization, which is represented in a neural network involving salience detection and self-processing with perceptual integration. As our society progresses in technological terms, having a greater conceptualization of how decision-making processes differ during interactions with humans and machines can provide pertinent information. A better understanding of those interactions can ultimately allow for safety measures to prevent any mishaps that can occur during advice taking.

### 3.6 References

- Ashburner, J., & Friston, K. J. (2005). Unified segmentation. *Neuroimage*, 26(3), 839-851.
- Beckmann, M., Johansen-Berg, H., & Rushworth, M. F. (2009). Connectivity-based parcellation of human cingulate cortex and its relation to functional specialization. *J Neurosci*, 29(4), 1175-1190.
- Benjamini, Y., & Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1), 289-300.
- Biele, G., Rieskamp, J., Krugel, L. K., & Heekeren, H. R. (2011). The Neural Basis of Following Advice. *PLoS Biol*, 9(6), e1001089.
- Boorman, E. D., O'Doherty, J. P., Adolphs, R., & Rangel, A. (2013). The behavioral and neural mechanisms underlying the tracking of expertise. *Neuron*, 80(6), 1558-1571.
- Botvinick, M. M. (2007). Conflict monitoring and decision making: reconciling two perspectives on anterior cingulate function. *Cogn Affect Behav Neurosci*, 7(4), 356-366.
- Breznitz, S. (2013). *Cry wolf: The psychology of false alarms*: Psychology Press.
- Buchel, C., Holmes, A. P., Rees, G., & Friston, K. J. (1998). Characterizing Stimulus-Response Functions Using Nonlinear Regressors in Parametric fMRI Experiments. *Neuroimage*, 8, 140-148.
- Chaminade, T., Rosset, D., Da Fonseca, D., Nazarian, B., Lutchter, E., Cheng, G., & Deruelle, C. (2012). How do we think machines think? An fMRI study of alleged competition with an artificial intelligence. *Frontiers in human neuroscience*, 6.
- Cooper, J. C., Kreps, T. A., Wiebe, T., Pirkel, T., & Knutson, B. (2010). When giving is good: ventromedial prefrontal cortex activation for others' intentions. *Neuron*, 67(3), 511-521.
- Costa, P., & McCrae, R. (1992). *Revised NEO Personality Inventory (NEO PI-R) and NEO Five-Factor Inventory (NEO-FFI) professional manual*. Odessa, FL: Psychological Assessment Resources, Inc.
- Crottaz-Herbette, S., & Menon, V. (2006). Where and when the anterior cingulate cortex modulates attentional response: combined fMRI and ERP evidence. *J Cogn Neurosci*, 18(5), 766-780.
- Davis, M. H. (1983). Measuring individual differences in empathy: Evidence for a multidimensional approach. *Journal of Personality and Social Psychology*, 44(1), 113-126.
- Delgado, M., Locke, H., Stenger, V., & Fiez, J. (2003). Dorsal striatum responses to reward and punishment: effects of valence and magnitude manipulations. *Cognitive, Affective, & Behavioral Neuroscience*, 3(1), 27-38.
- Dillon, D. G., Holmes, A. J., Jahn, A. L., Bogdan, R., Wald, L. L., & Pizzagalli, D. A. (2008). Dissociation of neural regions associated with anticipatory versus consummatory phases of incentive processing. *Psychophysiology*, 45(1), 36-49.

- Dixon, S. R., Wickens, C. D., & McCarley, J. S. (2007). On the independence of compliance and reliance: Are automation false alarms worse than misses? *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 49(4), 564-572.
- Dzindolet, M. T., Pierce, L. G., Beck, H. P., & Dawe, L. A. (2002). The perceived utility of human and automated aids in a visual detection task. *Human Factors*, 44(1), 79-94.
- Engelmann, J. B., Capra, C. M., Noussair, C., & Berns, G. S. (2009). Expert Financial Advice Neurobiologically “Offloads” Financial Decision-Making under Risk. *PloS one*, 4(3), e4957.
- Feng, C., Deshpande, G., Liu, C., Gu, R., Luo, Y.-J., & Krueger, F. (2015). Diffusion of responsibility attenuates altruistic punishment: a functional magnetic resonance imaging effective connectivity study. *Human Brain Mapping 2015, in press*.
- Forman, S. D., Cohen, J. D., Fitzgerald, M., Eddy, W. F., Mintun, M. A., & Noll, D. C. (1995). Improved assessment of significant activation in functional magnetic resonance imaging (fMRI): use of a cluster-size threshold. *Magn Reson Med*, 33(5), 636-647.
- Friston, K. J., Harrison, L., & Penny, W. (2003). Dynamic causal modelling. *Neuroimage*, 19(4), 1273-1302.
- Goebel, R., Esposito, F., & Formisano, E. (2006). Analysis of functional image analysis contest (FIAC) data with brainvoyager QX: From single-subject to cortically aligned group general linear model analysis and self-organizing group independent component analysis. *Hum Brain Mapp*, 27(5), 392-401.
- Goodyear, K., Parasuraman, R., Chernyak, S., Madhavan, P., Deshpande, G., & Krueger, F. (2015, submitted). Advice utilization during human and machine interactions: an fMRI and effective connectivity study. *Manuscript submitted for publication*.
- Grant, M. M., White, D., Hadley, J., Hutcheson, N., Shelton, R., Sreenivasan, K., & Deshpande, G. (2014). Early life trauma and directional brain connectivity within major depression. *Hum Brain Mapp*, 35(9), 4815-4826.
- Grant, M. M., Wood, K., Sreenivasan, K., Wheelock, M., & White, D. (2015). Influence of Early Life Stress on Intra- and Extra-Amygdaloid Causal Connectivity.
- Hampstead, B., Khoshnoodi, M., Yan, W., Deshpande, G., & Sathian, K. (2016). Patterns of effective connectivity during memory encoding and retrieval differ between patients with mild cognitive impairment and healthy older adults. *Neuroimage*, 124, 997-1008.
- Havlicek, M., Friston, K. J., Jan, J., Brazdil, M., & Calhoun, V. D. (2011). Dynamic modeling of neuronal responses in fMRI using cubature Kalman filtering. *Neuroimage*, 56(4), 2109-2128.
- Hutcheson, N. L., Sreenivasan, K. R., Deshpande, G., Reid, M. A., Hadley, J., White, D. M., . . . Lahti, A. C. (2015). Effective connectivity during episodic memory retrieval in schizophrenia participants before and after antipsychotic medication. *Hum Brain Mapp*, 36(4), 1442-1457.

- Krach, S., Hegel, F., Wrede, B., Sagerer, G., Binkofski, F., & Kircher, T. (2008). Can Machines Think? Interaction and Perspective Taking with Robots Investigated via fMRI. *PloS one*, 3(7), e2597.
- Kriegeskorte, N., Simmons, W. K., Bellgowan, P. S. F., & Baker, C. I. (2009). Circular analysis in systems neuroscience: the dangers of double dipping. *Nat Neurosci*, 12(5), 535-540.
- Lacey, S., Stilla, R., Sreenivasan, K., Deshpande, G., & Sathian, K. (2014). Spatial imagery in haptic shape perception. *Neuropsychologia*, 60, 144-158.
- Lewandowsky, S., Mundy, M., & Tan, G. (2000). The dynamics of trust: comparing humans to automation. *Journal of Experimental Psychology: Applied*, 6(2), 104.
- Litt, A., Plassmann, H., Shiv, B., & Rangel, A. (2011). Dissociating valuation and saliency signals during decision-making. *Cereb Cortex*, 21(1), 95-102.
- Madhavan, P., & Gonzalez, C. (2006). *Effects of sensitivity, criterion shifts, and subjective confidence on the development of automaticity in airline luggage screening*. Paper presented at the Proceedings of the Human Factors and Ergonomics Society Annual Meeting.
- Madhavan, P., & Wiegmann, D. A. (2007). Effects of Information Source, Pedigree, and Reliability on Operator Interaction With Decision Support Systems. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 49(5), 773-785.
- McBride, S. E., Rogers, W. A., & Fisk, A. D. (2014). Understanding human management of automation errors. *Theoretical issues in ergonomics science*, 15(6), 545-577.
- Merritt, S. M., Heimbaugh, H., LaChapell, J., & Lee, D. (2013). I Trust It, But I Don't Know Why : Effects of Implicit Attitudes Toward Automation on Trust in an Automated System. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 55(3), 520-534.
- Meshi, D., Biele, G., Korn, C. W., & Heekeren, H. R. (2012). How expert advice influences decision making. *PloS one*, 7(11), e49748.
- Mesulam, M. M. (1998). From sensation to cognition. *Brain*, 121 ( Pt 6), 1013-1052.
- Onnasch, L., Ruff, S., & Manzey, D. (2014). Operators' adaptation to imperfect automation – Impact of miss-prone alarm systems on attention allocation and performance. *International Journal of Human-Computer Studies*, 72(10–11), 772-782.
- Packard, M. G., & Knowlton, B. J. (2002). Learning and memory functions of the basal ganglia. *Annual review of neuroscience*, 25(1), 563-593.
- Parasuraman, A. (2000). Technology Readiness Index (Tri): A Multiple-Item Scale to Measure Readiness to Embrace New Technologies. *Journal of Service Research*, 2(4), 307-320.
- Rice, S., & McCarley, J. S. (2011). Effects of Response Bias and Judgment Framing on Operator Use of an Automated Aid in a Target Detection Task. *Journal of Experimental Psychology: Applied*, 17(4), 320-331.
- Schilbach, L., Eickhoff, S. B., Schultze, T., Mojzisch, A., & Vogeley, K. (2013). To you I am listening: perceived competence of advisors influences judgment and decision-making via recruitment of the amygdala. *Soc Neurosci*, 8(3), 189-202.

- Singh, I. L., Molloy, R., & Parasuraman, R. (1997). Automation-induced "complacency": development of the complacency-potential rating scale. *The International Journal of Aviation Psychology*, 3(2), 111-122.
- Suen, V. Y. M., Brown, M. R. G., Morck, R. K., & Silverstone, P. H. (2014). Regional Brain Changes Occurring during Disobedience to "Experts" in Financial Decision-Making. *PloS one*, 9(1), e87321.
- Terasawa, Y., Fukushima, H., & Umeda, S. (2013). How does interoceptive awareness interact with the subjective experience of emotion? An fMRI study. *Hum Brain Mapp*, 34(3), 598-612.
- Van Swol, L. M., & Snizek, J. A. (2005). Factors affecting the acceptance of expert advice. *Br J Soc Psychol*, 44(Pt 3), 443-461.
- Vickery, T. J., & Jiang, Y. V. (2009). Inferior parietal lobule supports decision making under uncertainty in humans. *Cereb Cortex*, 19(4), 916-925.
- Wang, Y., & Quadflieg, S. (2015). In our own image? Emotional and neural processing differences when observing human-human vs human-robot interactions. *Soc Cogn Affect Neurosci*.
- Wheelock, M. D., Sreenivasan, K. R., Wood, K. H., Ver Hoef, L. W., Deshpande, G., & Knight, D. C. (2014). Threat-related learning relies on distinct dorsal prefrontal cortex network connectivity. *Neuroimage*, 102 Pt 2, 904-912.
- Xia, M., Wang, J., & He, Y. (2013). BrainNet Viewer: A Network Visualization Tool for Human Brain Connectomics. *PloS one*, 8(7), e68910.

## **CHAPTER FOUR: GENERAL DISCUSSION**

This thesis has examined the impact of misses and false alarms during advice utilization from human and machine agents in a series of two studies. The goal of this thesis was to provide a basis for understanding the complex neural and behavioral mechanisms involved during advice utilization, which can ultimately serve to develop a framework underlining the constituents of human and machine interactions. In each study we demonstrated that there were unique behavioral responses and brain activation patterns associated with each error type. The rest of Chapter Four will generally discuss the behavioral and brain activation differences across the studies along with future directions.

### **4.1 Behavioral Results**

In Chapter Two and Chapter Three we revealed that the no agent groups performed significantly better than the human and machine agent groups. The results indicate that regardless of error, individuals who performed the task unassisted, and did not receive unreliable advice, performed better overall. It has been postulated that false alarms cause individuals to ignore true alerts leading to a decline in performance, while misses create higher workloads from increased monitoring, which also affects performance (Sanchez, Rogers, Fisk, & Rovira, 2014). We therefore expected in both

studies, that unreliable advice would decrease performance due to high error rates and that participants in the no agent groups would perform significantly better. Despite the fact that participant performance in the no agent groups was higher than performance in both studies, the accuracy rate was still not ideal for any real world applications. These results align with the findings by Wickens and Dixon (2007) that showed that automation reliability below 70% significantly decreased performance compared to performing the task unassisted. Although our study included both humans and machines with low reliability, it is possible that the optimal reliability set point is not necessarily dependent on the source of advice. Moreover, our results provide evidence that misses could have created higher vigilance in performance compared to false alarms, which may have fewer repercussions if ignored.

In Chapter Two we showed that advice utilization degraded more for the human-agent group, while in Chapter Three advice utilization degraded more for the machine-agent group. We hypothesized that advice utilization would decrease more for the machine-agent group in both studies due to previous findings that showed that when advice is 70% reliable, participants agree more with expert humans and depend less on expert machines (Madhavan & Wiegmann, 2007). However, the study by Madhavan and Wiegmann (2007) focused on the combination of false alarms and misses without separating the two error types and that might explain why we found differences in Chapter Two compared to Chapter Three. Our results further indicate that accountability may be higher during interactions with a human when an error is a false alarm and when an error is a miss, accountability may be higher during interactions with a machine.

Previous work has revealed that 70% reliable automation may disrupt preconceived notions associated with the perfect automation schema due to the effects of dispositional factors associated with advisors (Madhavan & Wiegmann, 2007) and participant's perceived accountability for their performance may be due to automation bias, or the tendency toward usage of, or reliance on, automation without actively seeking or processing information (Mosier et al., 1998). Our results reflect a disruption in the perfect automation schema, or biases associated with automation when the error was a miss, which could be due to the costly consequences of a miss error.

For reliability, we revealed in Chapter Two that the human agent's pre-reliability was significantly higher than the machine agent's pre-reliability and the reliability ratings significantly decreased pre- to post-experiment for the human-agent group. Furthermore, the human agent's perceived reliability was significantly higher than the actual reliability of the agent. These results suggest that expectations of reliable advice were higher for the human-agent group compared to the machine-agent group, which ultimately led to a behavioral adjustment in advice utilization over time. In comparison, for Chapter Three, the reliability ratings did not differ between the agent groups, but the perceived reliability ratings for both the human agent and machine agent were significantly higher than the actual reliability, showing that initial expectations of reliable advice were high for both groups. Initial expectations of reliable advice, as seen during the comparison of the perceived reliability to the actual reliability of each agent, can lead to a decline in dependence on an agent and miscalibration of an agent's reliability (Madhavan & Wiegmann, 2007). The reliability ratings were initially higher than the actual reliability

for the human agent in Chapter Two and for both groups in Chapter Three, indicating high expectations of reliable advice. However, upon observation of the errors (40%) generated by the agents, the participant's advice utilization degraded rapidly. Moreover, in Chapter Two, the machine agent's perceived pre-reliability ratings were not significantly different from the actual reliability of the agent, showing that initial expectations of reliability were not high and thus participants may not have needed to recalibrate their expectations as indicated by less degradation of advice utilization.

In Chapter Two, we demonstrated that trust significantly decreased for the human agent, however in Chapter Three, trust decreased for both groups. In Chapter Two, advice utilization decreased more for the human-agent group compared to the machine-agent group, which was also reflected by the change in trust ratings only for the human-agent group. Similarly, in Chapter Three, advice utilization decreased for both groups, which was also reflected in the change in trust ratings for both groups. It has been suggested that user attitudes such as trust may affect how individuals decide to use automation (Lee & See, 2004). For example, a study showed that human experts were trusted more than machine experts (Madhavan & Wiegmann, 2007) which indicates that trust may be one of the components involved during advice utilization interactions for both humans and machines.

Lastly, we looked at confidence ratings and for Chapter Two and Chapter Three we showed that confidence was rated higher on target bags compared to non-target bags. However, we showed no difference between the agent groups for confidence ratings. Previous research has indicated that self-confidence may affect decision biases, which

may change performance accuracy (Madhavan & Gonzalez, 2006) and when trust exceeds self-confidence, individuals tend towards automation use (Lee & Moray, 1992). Since our findings did not show differences between the agent groups, the differences in advice utilization between the human and machine agents cannot be explained by self-confidence.

For response times, we found that for both Chapter Two and Chapter Three responses were faster during run 2 compared to run 1 and for Chapter Two, responses were faster during good advice compared to bad advice. These results indicate that as participants became more familiar with the task they were able to respond faster to the advice given. Furthermore, participants in Chapter Two may have had more conflicting perceptual processes involved during false alarm trials as reflected by slower responses during bad advice. Research on response times have demonstrated that false alarms may result in a delayed or no response to alerts (Breznitz, 2013) and our results are in accordance with those findings.

Monetary deductions were used as incentives and as a way to create a risky environment for participants in order to help evaluate variables such as trust towards the human and machine agents. In Chapter Two, we found that deductions were higher during bad advice compared to good advice; in Chapter Three we found that deductions were higher during run 1 compared to run 2. The impact of errors on monetary deductions was revealed as participants made more costly errors in Chapter Two, while in Chapter Three, participants made less costly errors over time.

## 4.2 FMRI Results

In Chapter Two, we revealed a network that involved brain regions associated with social evaluations (aPreC, PCC), while in Chapter Three there was a network engaged with visual processing of sensory information (LG). As expected, the comparison of the studies show that there are distinct neural networks involved with false alarms compared to misses during advice utilization from human and machine agents. Our results are in line with the findings of Onnasch et al. (2014) and Breznitz (2013), that false alarms may cause operators to have delayed responses, or no response at all, while misses may change operator's strategies during non-alarm periods causing a reallocation of attention. In Chapter Two we revealed a brain network involved with social evaluations of the dispositional characteristics of the agents, while in Chapter Three there was a network involved with visual processing and error monitoring, as participants shifted their attention towards the task at hand. Since false alarms are not necessarily detrimental, but more of a nuisance, participants may have had more time to evaluate human traits such as trust or agent effort leading to involvement of regions associated with social evaluations. On the other hand, due to the catastrophic nature of misses, participants may have concentrated more on situational factors, such as task difficulty, which was reflected by recruitment of visual processing regions. The comparisons between Chapter Two and Chapter Three during the decision phase provides evidence that there are separate perceptual processes involved with each error type, which has also been demonstrated with changes in cortical activity during a contrast-detection task comparing misses to false alarms (Ress & Heeger, 2003).

Interestingly, the feedback phase results demonstrated a similar pattern to that of the decision phase results for both studies, with areas involved with social evaluations (dmPFC) and processing of sensory information (FG, IPL). The results indicate that there was a unique pattern of activity for brain regions involved during the feedback phase as participants were able to evaluate their own performance based on the advice given to them. These findings are of particular importance because it provides a greater discernment of the underlying mechanisms involved during learning and behavioral adaptations to unreliable advice. As with the decision phase, the feedback phase results for Chapter Two and Chapter Three provides evidence that there may be distinct processes involved with perceptions of different error types.

### **4.3 Future Directions and Conclusions**

The findings of Chapter Two and Chapter Three provide insight into the differences between error types during decision-making, which ultimately serves to optimize our understanding of how individuals choose to utilize or discount advice from different agents. Future studies could elaborate on our findings by implementing a paradigm with agent reliability above the 70% threshold to investigate the behavioral responses and the underlying brain network involved with reliable advice. Furthermore, future studies could expand on our results by implementing a paradigm with no feedback, or positive and negative feedback, mirroring human etiquette. Additionally, we aimed to discern the effective connectivity network associated with advice utilization with Granger Causality Analysis. Granger Causality Analysis was used for our studies since it is

particularly advantageous for exploratory analysis and for assessing directional influences of selected ROIs without an *a priori* hypothesis. To further validate our findings, future studies could implement methods such as dynamic causal modeling (DCM) with a hypothesized network that is predefined to model the effective connectivity results that we discovered.

In conclusion, this thesis has aimed to uncover the factors that influence advice utilization from humans and machines by assessing the behavioral responses and neural mechanisms associated with those interactions. The overall objective of this research was to provide a foundation that will facilitate the development of a cohesive model explaining the behavioral, cognitive, and neural basis of advice utilization during human-automation interactions by bridging the gap between human factors and cognitive neuroscience research. The findings of this thesis are especially salient for the future as technological progressions continue to increase exponentially and the shift to automation use becomes inevitable.

## APPENDIX A: FALSE ALARMS

### A.1 Experimental Setup

a)



X-ray Bag

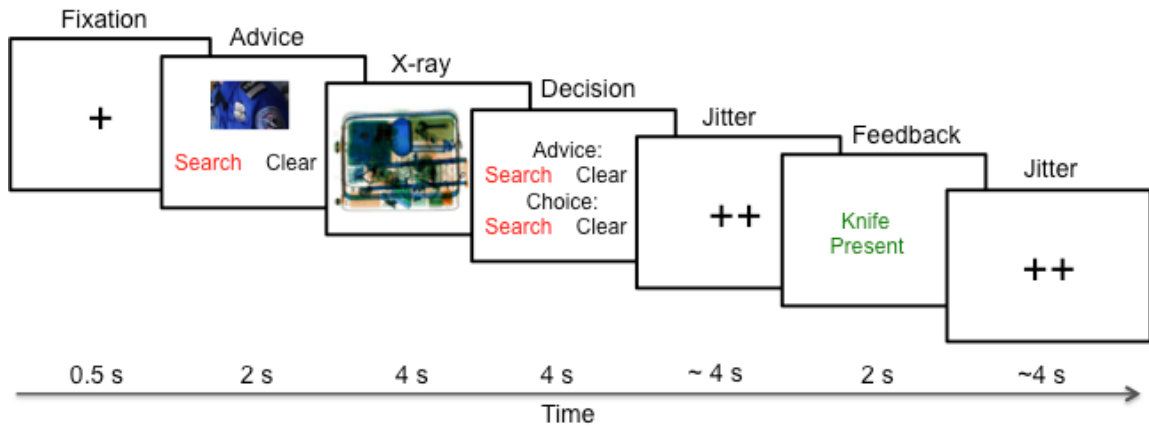


Target

b)

		Advice			
		Yes	No		
Target	Yes	Hit (50%)	Miss (0%)	Good Advice	
	No	False Alarm (40%)	Correct Rejection (10%)	Bad Advice	

c)



**A.1. a) Example Stimuli Used for the X-ray Luggage-screening Task.** During the normative rating task, participants rated 320 X-ray luggage images (120 target: 60 high clutter, 60 low clutter; 200 non-target: 100 high clutter, 100 low clutter) that contained everyday objects (hair-dryers, clothes, etc.) and a possible target present (5 different knives, with one possible per image) based on clutter, difficulty and confidence in finding the knife. **b) Decision Matrix.** Breakdown for each advice type given during the experiment. **c) X-ray Luggage-Screening Task.** During each trial, participants would first see a fixation cross, advice from one of the agents to “search” or “clear” the bag, an image of the X-ray luggage bag, a decision to accept or reject the advice of the agent to “search” or “clear” the bag, fixation crosses, feedback indicating if their decision was correct or incorrect and lastly, fixation crosses.

## **A.2 Human and Machine Agent Descriptions**

### **Human: Mr. Steve Williams**

Mr. Steve Williams (Human) is a trained luggage screener, with extensive knowledge in identifying illegal imports inside airline luggage. He has served the past 5 years in some of the busiest airports in the United States working at security checkpoints. He also specializes in antiterrorism and airport security and possesses extensive knowledge about the types of modern weapons and explosives commonly smuggled aboard aircraft. Mr. Williams has recently been appointed by the Transportation Security Administration (TSA) to oversee security operations at Dulles International Airport, which is one of the largest airports in the world.

### **Machine: Automated Luggage Inspector**

The automated luggage inspector (Machine) is a diagnostic aid that has been programmed to identify hidden contraband in airline luggage. This Machine is based upon the technology traditionally used at major airport security checkpoints over the past 5 years. Its algorithms are sophisticated and are based on judgments using sensors different from those of the human visual system and can detect modern weapons and explosives smuggled aboard aircrafts. The automated luggage detector has recently been employed by the Transportation Security Administration (TSA) to enhance security operations at Dulles International Airport, which is one of the largest airports in the world.

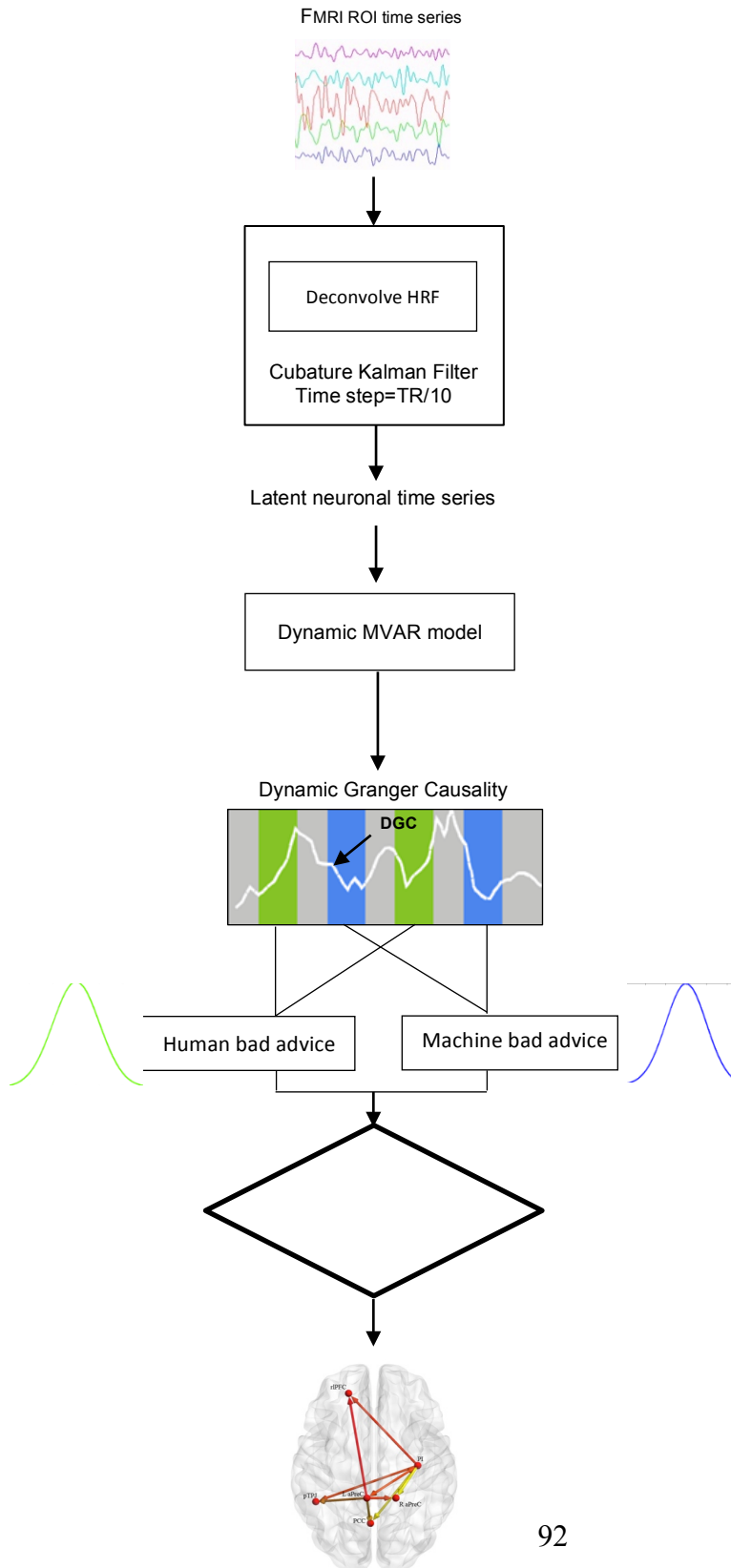
### A.3 Brain Regions Associated with the Main Effect of Advice

Brain regions showing significant activation clusters associated during the decision

(minimum cluster of 21) and feedback (minimum cluster of 36) phases ( $\alpha < .05$ , cluster-level threshold corrected). For the decision phase, a significant activation cluster was found in the right orbitofrontal cortex (superior frontal gyrus, BA 11). For the feedback phase, significant activation clusters were found in right middle frontal gyrus (BA 6/8), right superior parietal lobule (BA 7), right putamen, right posterior cingulate cortex (BA 30), right head of the caudate, left orbitofrontal cortex (medial frontal gyrus, BA 11), left precentral gyrus (BA 4), left subcallosal gyrus (BA 34), left middle frontal gyrus (BA 6), left dorsolateral prefrontal cortex (middle frontal gyrus, BA 46) and left inferior frontal gyrus (BA 47).

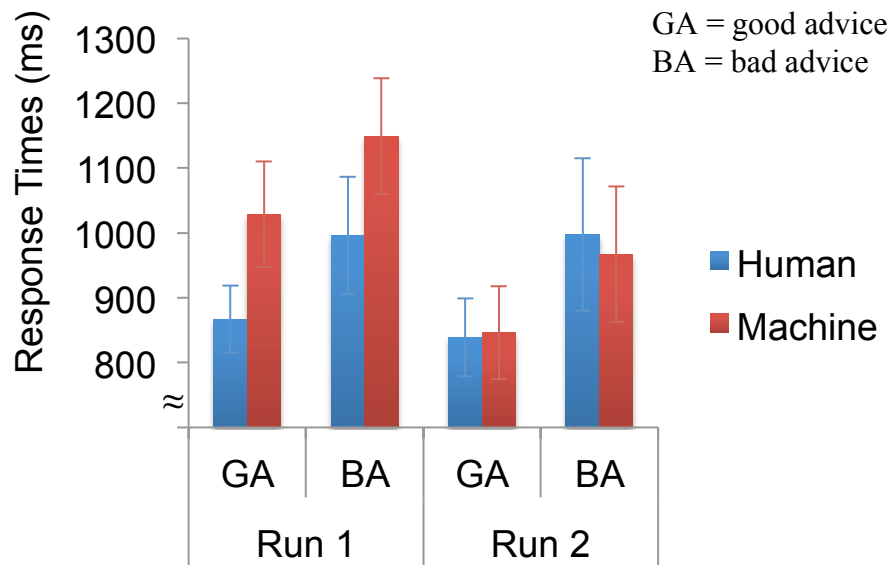
	<i>F</i> (1,22) value	Cluster Size (mm <sup>3</sup> )	x	y	z
<b>Decision Phase</b>					
<i>Advice</i>					
Right orbitofrontal cortex	13.14	673	18	45	-18
<b>Feedback Phase</b>					
<i>Advice</i>					
Right middle frontal gyrus	16.47	4848	36	18	57
Right superior parietal lobule	13.05	2010	21	-45	57
Right putamen	12.18	1867	33	-3	3
Right posterior cingulate cortex	12.47	4937	6	-51	15
Right head of the caudate	14.27	1968	9	12	-9
Left orbitofrontal cortex	12.30	3348	-9	48	-15
Left precentral gyrus	15.29	4486	-24	-24	63
Left subcallosal gyrus	13.08	2204	-12	3	-12
Left middle frontal gyrus	12.05	2553	-33	25	60
Left dorsolateral prefrontal cortex	15.05	2228	-42	36	12
Left inferior frontal gyrus	15.75	1778	-36	27	-6

## A.4 Schematic Illustrating the Effective Connectivity Analysis Pipeline

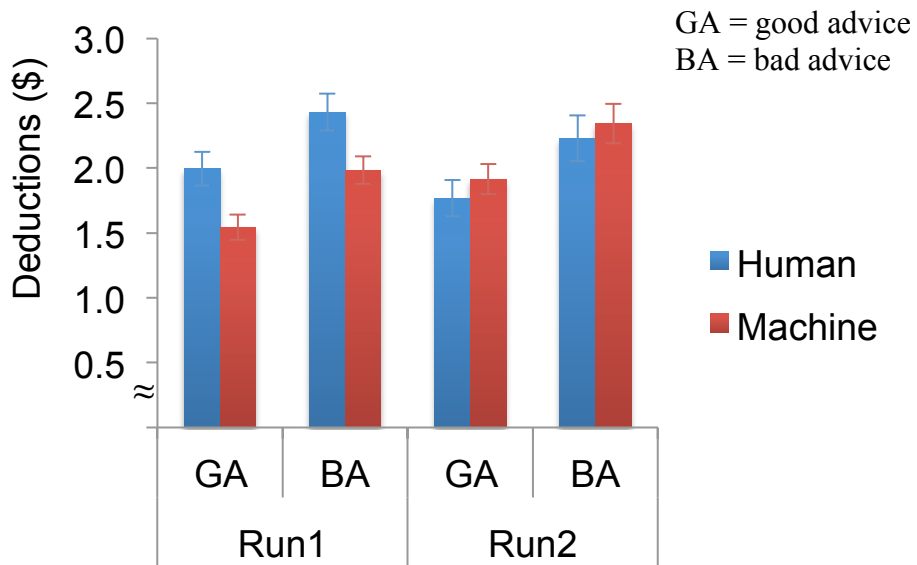


## A.5 Behavioral Results for Decision Phase

a)

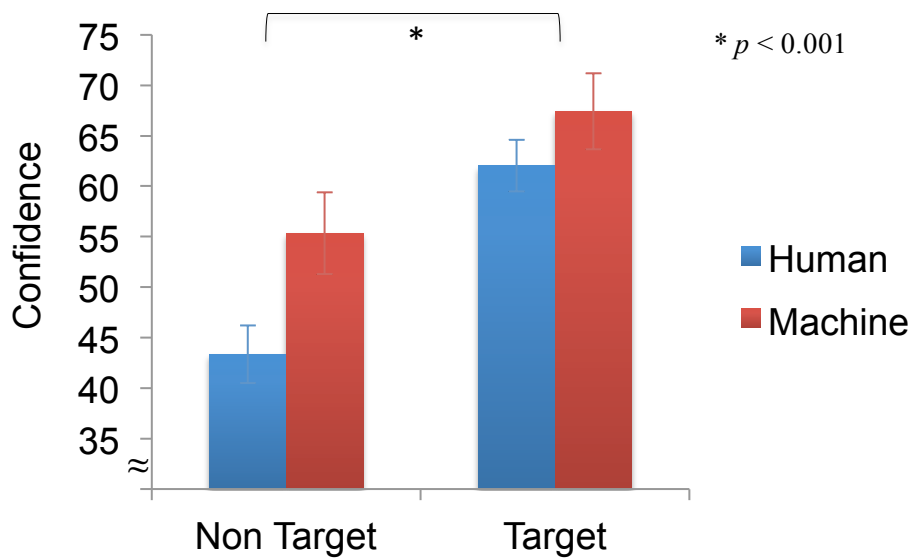


b)



**A.5. ( $M \pm SEM$ ). a) Response Times.** Response times were faster overall from run 1 to run 2 and during good advice compared to bad advice. **b) Monetary Deductions.** Monetary deductions were higher overall for bad advice compared to good advice.

#### A.6 Results for the Confidence Ratings



**A.6. ( $M \pm SEM$ ).** Confidence ratings were significantly lower during non-target bags compared to target bags.

## A.7 Descriptive Statistics for Psychological Control Measures

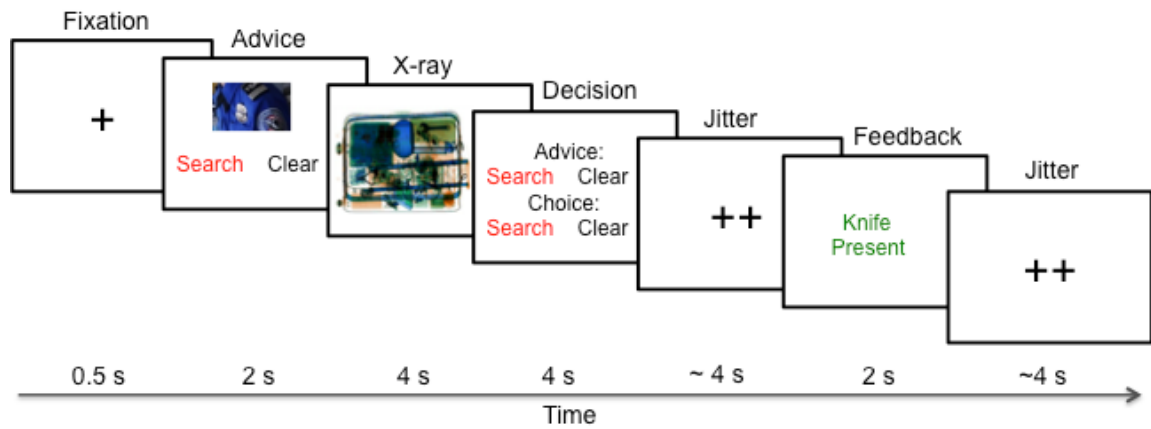
No significant differences were found between the human- and machine-agent groups ( $M \pm SD$ ).

Category	Human	Machine	Statistics
<b>Demographics</b> <b>df = 22</b>			
Age	20.33 $\pm$ 2.55	20.42 $\pm$ 2.75	$t = -0.08, p = .939$
Education	14.08 $\pm$ 2.35	14.13 $\pm$ 1.65	$t = -0.50, p = .960$
Handedness	96.53 $\pm$ 8.31	92.49 $\pm$ 6.77	$t = 1.31, p = .205$
Gender (male/female)	7/5	6/6	$\chi^2 = 0.17, p = .683$
<b>Complacency-Potential Rating Scale (CPS)</b>			
Confidence	15.17 $\pm$ 2.13	14.50 $\pm$ 1.78	$t = 0.83, p = .414$
Reliance	9.50 $\pm$ 1.68	10.33 $\pm$ 1.78	$t = -1.18, p = .250$
Trust	8.58 $\pm$ 2.28	8.92 $\pm$ 1.44	$t = -0.43, p = .672$
Safety	6.25 $\pm$ 1.71	6.75 $\pm$ 2.09	$t = -0.64, p = .529$
<b>Interpersonal Reactivity Index (IRI)</b>			
Perspective Taking	28.25 $\pm$ 2.30	28.33 $\pm$ 3.37	$t = -0.71, p = .944$
Fantasy Scale	19.33 $\pm$ 2.84	20.25 $\pm$ 2.80	$t = -0.80, p = .434$
Empathic Concern	21.67 $\pm$ 5.07	22.33 $\pm$ 2.39	$t = -0.41, p = .684$
Personal Distress	20.75 $\pm$ 2.80	20.67 $\pm$ 2.96	$t = 0.71, p = .944$
<b>NEO Five-Factor Inventory (NEO-FFI)</b>			
Neuroticism	31.33 $\pm$ 4.89	32.67 $\pm$ 3.94	$t = -0.74, p = .470$
Extraversion	41.92 $\pm$ 3.37	40.42 $\pm$ 3.26	$t = 1.11, p = .280$
Openness	37.75 $\pm$ 3.60	36.92 $\pm$ 4.72	$t = 0.49, p = .631$
Agreeableness	38.67 $\pm$ 4.05	41.00 $\pm$ 4.51	$t = -1.33, p = .196$
Conscientiousness	41.50 $\pm$ 3.56	42.17 $\pm$ 3.49	$t = -0.46, p = .647$
<b>National Technology Readiness Survey (NTRS)</b>			
Optimism	37.58 $\pm$ 4.87	39.08 $\pm$ 4.54	$t = -0.78, p = .444$
Innovativeness	21.75 $\pm$ 4.20	24.83 $\pm$ 4.24	$t = -1.79, p = .087$
Discomfort	31.00 $\pm$ 5.21	31.50 $\pm$ 5.02	$t = -0.24, p = .813$
Insecurity	30.33 $\pm$ 5.68	29.08 $\pm$ 3.85	$t = 0.63, p = .534$
<b>Propensity to Trust (PTT)</b>			
Trust towards Automation	19.83 $\pm$ 2.21	20.42 $\pm$ 2.07	$t = -0.67, p = .511$

## APPENDIX B: MISSES

### B.1 Experimental Setup

a)



b)

		Advice	
		Yes	No
Target	Yes	Hit (10%)	Miss (40%)
	No	False Alarm (0%)	Correct Rejection (50%)

Good Advice

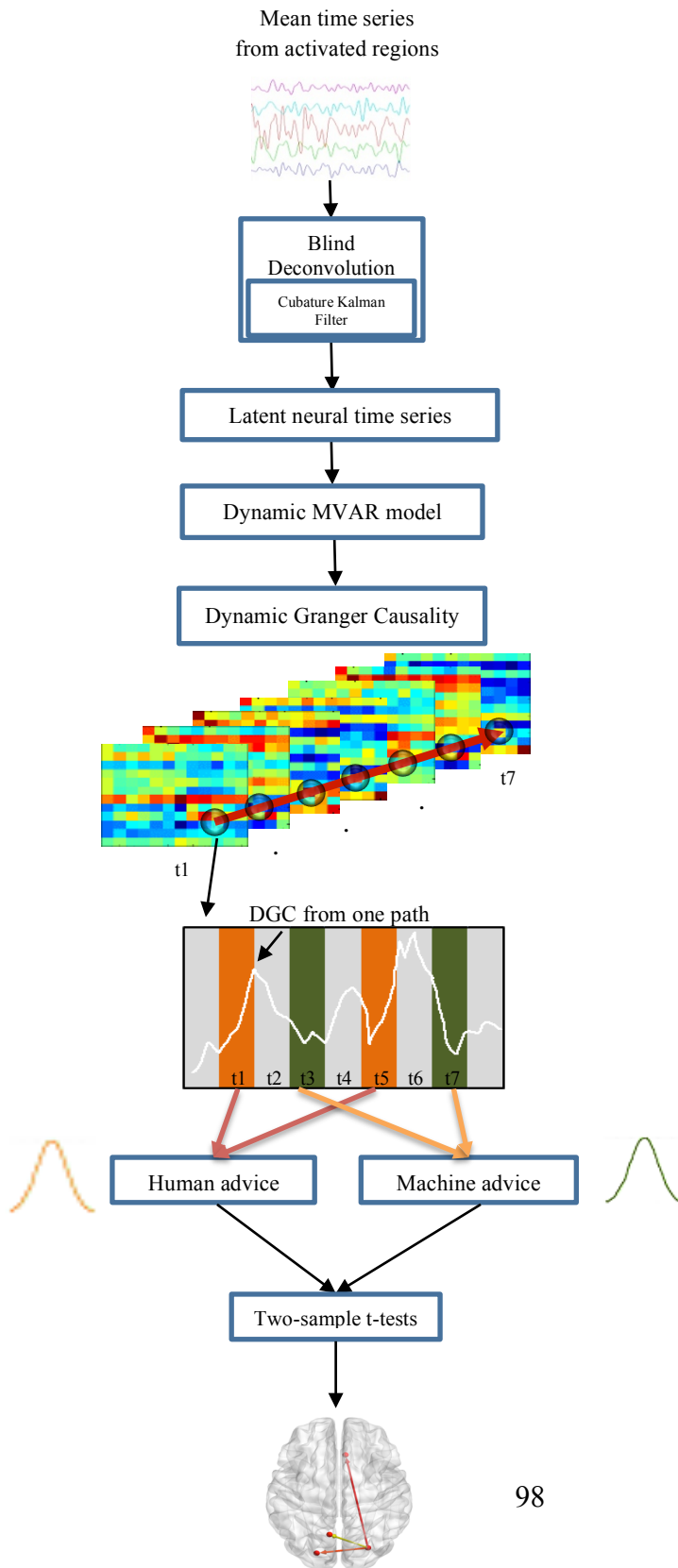
Bad Advice

**B.1. a) X-ray Luggage-Screening Task.** During each trial, participants would first see a fixation cross, advice from one of the agents to “search” or “clear” the bag, an image of the X-ray luggage bag, a decision to accept or reject the advice of the agent to “search” or “clear” the bag, fixation crosses, feedback indicating if their decision was correct or incorrect and lastly, fixation crosses. **b) Decision Matrix.** Breakdown for each advice type (good, bad) given during the experiment.

## **B.2 Effective Connectivity Analysis**

Granger causality is based on a concept of causality that can be used to predict directional influences among chosen brain regions through multivariate effective connectivity modeling of ROI (region of interest) time courses (Deshpande, LaConte, James, Peltier, & Hu, 2009; Friston, Harrison, & Penny, 2003; Granger, 1969; Preusse, van der Meer, Deshpande, Krueger, & Wartenburger, 2011). The model examines the relationship of variables in time, such that given two variables,  $a$  and  $b$ , if past values of  $a$  better predict the present value of  $b$ , then as a function of earlier time points, causality between the variables can be inferred (Goodyear et al., 2015, submitted; Hampstead et al., 2011; Krueger, Landgraf, van der Meer, Deshpande, & Hu, 2011; Roebroek, Formisano, & Goebel, 2005). Granger causality analysis is a data-driven approach and thus is advantageous for application of effective connectivity since there is no requirement for pre-specified connectivity models like dynamic causal modeling (Deshpande & Hu, 2012; Deshpande et al., 2009; Deshpande, Sathian, & Hu, 2010).

### B.3 Schematic Illustrating the Effective Connectivity Analysis Pipeline.



**B.3.** The mean time series from the ROIs from the decision and feedback phases were extracted, then blind hemodynamic deconvolution was performed using a Cubature Kalman Filter to reveal the underlying latent neural time series. Next, these time series were applied to a dynamic Multivariate Autoregressive Model based on a Granger causality framework. Granger connectivity path weights were populated into two samples and t-tests were performed for each effective connectivity path to reveal those that were significantly different between the agent groups.

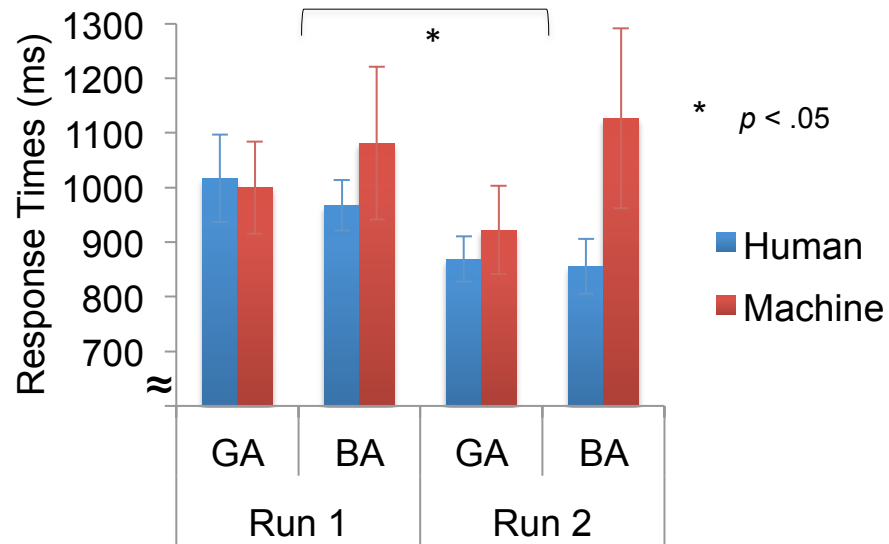
## B.4 Descriptive Statistics for Psychological Control Measures

No significant differences were found between the human- and machine-agent groups ( $M \pm SD$ ).

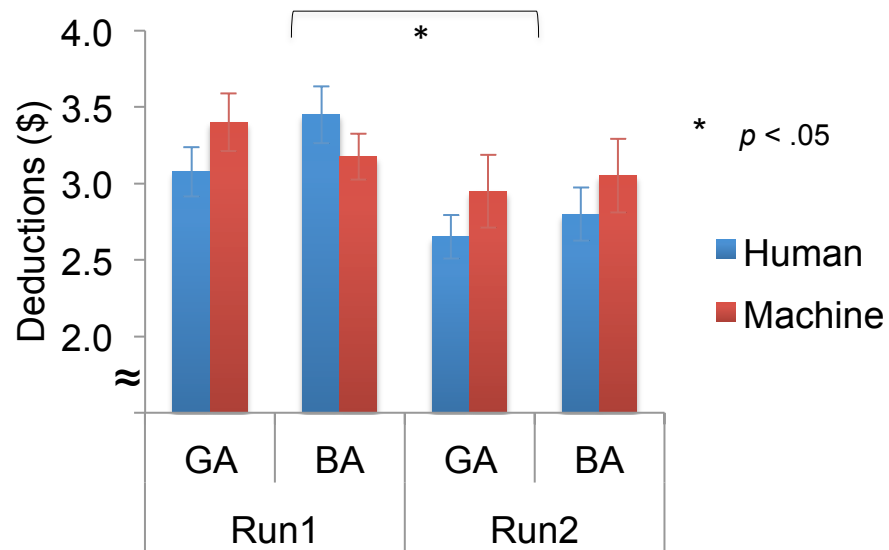
Category	Human	Machine	Statistics
<b>Demographics df = 22</b>			
Age	22.58 $\pm$ 2.39	21.92 $\pm$ 2.43	$t = 0.68, p = .505$
Education	16.25 $\pm$ 1.71	16.08 $\pm$ 2.68	$t = 0.18, p = .858$
Gender (male/female)	7/5	7/5	$\chi^2(1) = 0.67, p = .414$
<b>Complacency-Potential Rating Scale (CPS; feelings toward automation)</b>			
Confidence	16.17 $\pm$ 2.41	15.42 $\pm$ 1.78	$t = 0.89, p = .395$
Reliance	10.50 $\pm$ 1.51	10.08 $\pm$ 1.44	$t = 0.69, p = .496$
Trust	10.17 $\pm$ 1.95	8.67 $\pm$ 1.92	$t = 1.90, p = .071$
Safety	6.50 $\pm$ 1.68	6.00 $\pm$ 1.35	$t = 0.80, p = .430$
<b>Interpersonal Reactivity Index (IRI; separate facet of empathy)</b>			
Perspective Taking	27.83 $\pm$ 2.13	26.58 $\pm$ 3.29	$t = 1.11, p = .281$
Fantasy Scale	19.00 $\pm$ 3.72	20.58 $\pm$ 1.51	$t = -1.37, p = .185$
Empathic Concern	22.83 $\pm$ 2.25	22.92 $\pm$ 2.19	$t = -0.09, p = .928$
Personal Distress	19.92 $\pm$ 3.06	19.33 $\pm$ 2.27	$t = 0.53, p = .601$
<b>NEO Five-Factor Inventory (NEO-FFI; personality styles)</b>			
Neuroticism	31.83 $\pm$ 3.56	33.50 $\pm$ 3.83	$t = -1.10, p = .281$
Extraversion	41.25 $\pm$ 4.69	40.83 $\pm$ 3.71	$t = 0.24, p = .812$
Openness	36.50 $\pm$ 4.44	35.83 $\pm$ 2.86	$t = 0.43, p = .666$
Agreeableness	38.17 $\pm$ 4.45	37.50 $\pm$ 4.98	$t = 0.35, p = .733$
Conscientiousness	43.25 $\pm$ 3.08	42.17 $\pm$ 4.32	$t = 0.71, p = .487$
<b>National Technology Readiness Survey (NTRS; embracing new technologies)</b>			
Optimism	38.50 $\pm$ 4.72	37.75 $\pm$ 5.97	$t = 0.34, p = .736$
Innovativeness	20.92 $\pm$ 5.62	22.58 $\pm$ 4.10	$t = -0.83, p = .415$
Discomfort	28.25 $\pm$ 5.64	30.50 $\pm$ 4.85	$t = -1.05, p = .306$
Insecurity	28.67 $\pm$ 5.09	28.50 $\pm$ 3.85	$t = 0.09, p = .929$
<b>Propensity to Trust (PTT; trust towards automation)</b>			
Trust towards Automation	21.17 $\pm$ 2.04	21.33 $\pm$ 2.10	$t = -0.20, p = .846$

## B.5 Results for the Decision Phase

a)

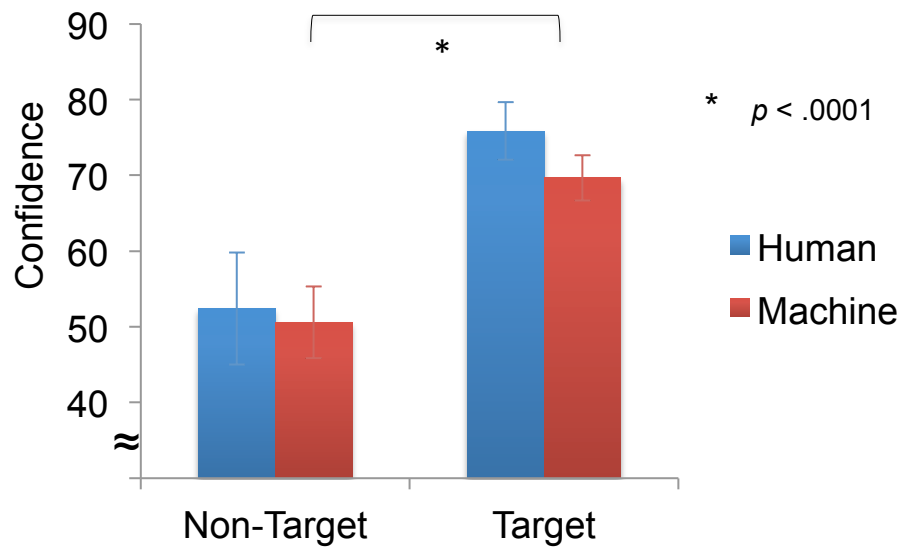


b)



**B.5. ( $M \pm SEM$ ).** a) **Response Times.** Response times were faster overall during run 2 compared to run 1. b) **Monetary Deductions.** Monetary deductions were higher during run 1 compared to run 2. GA = good advice; BA = bad advice.

## B.6 Confidence Ratings Results



**B.6. ( $M \pm SEM$ ).** Confidence ratings were significantly lower during non-target bags compared to target bags.

## B.7 Appendix B References

- Deshpande, G., & Hu, X. (2012). Investigating effective brain connectivity from fMRI data: past findings and current issues with reference to Granger causality analysis. *Brain Connect*, 2(5), 235-245.
- Deshpande, G., LaConte, S., James, G. A., Peltier, S., & Hu, X. (2009). Multivariate Granger causality analysis of fMRI data. *Hum Brain Mapp*, 30(4), 1361-1373.
- Deshpande, G., Sathian, K., & Hu, X. (2010). Effect of hemodynamic variability on Granger causality analysis of fMRI. *Neuroimage*, 52(3), 884-896.
- Friston, K. J., Harrison, L., & Penny, W. (2003). Dynamic causal modelling. *Neuroimage*, 19(4), 1273-1302.
- Goodyear, K., Parasuraman, R., Chernyak, S., Madhavan, P., Deshpande, G., & Krueger, F. (2015, submitted). Advice utilization during human and machine interactions: an fMRI and effective connectivity study. *Manuscript submitted for publication*.
- Granger, C. W. (1969). Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: Journal of the Econometric Society*, 424-438.
- Hampstead, B. M., Stringer, A. Y., Stilla, R. F., Deshpande, G., Hu, X., Moore, A. B., & Sathian, K. (2011). Activation and effective connectivity changes following explicit-memory training for face-name pairs in patients with mild cognitive impairment: a pilot study. *Neurorehabil Neural Repair*, 25(3), 210-222.
- Krueger, F., Landgraf, S., van der Meer, E., Deshpande, G., & Hu, X. (2011). Effective connectivity of the multiplication network: a functional MRI and multivariate Granger Causality Mapping study. *Hum Brain Mapp*, 32(9), 1419-1431.
- Preusse, F., van der Meer, E., Deshpande, G., Krueger, F., & Wartenburger, I. (2011). Fluid intelligence allows flexible recruitment of the parieto-frontal network in analogical reasoning. *Front Hum Neurosci*, 5, 22.
- Roebroeck, A., Formisano, E., & Goebel, R. (2005). Mapping directed influence over the brain using Granger causality and fMRI. *Neuroimage*, 25(1), 230-242.

## REFERENCES

- Boorman, E. D., O'Doherty, J. P., Adolphs, R., & Rangel, A. (2013). The behavioral and neural mechanisms underlying the tracking of expertise. *Neuron*, 80(6), 1558-1571.
- Breznitz, S. (2013). *Cry wolf: The psychology of false alarms*: Psychology Press.
- Brosch, T., Schiller, D., Mojdehbakhsh, R., Uleman, J. S., & Phelps, E. A. (2013). Neural mechanisms underlying the integration of situational information into attribution outcomes. *Soc Cogn Affect Neurosci*, 8(6), 640-646.
- Bush, G., Luu, P., & Posner, M. I. (2000). Cognitive and emotional influences in anterior cingulate cortex. *Trends in cognitive sciences*, 4(6), 215-222.
- Cabanis, M., Pyka, M., Mehl, S., Muller, B. W., Loos-Jankowiak, S., Winterer, G., . . . Kircher, T. (2013). The precuneus and the insula in self-attributional processes. *Cogn Affect Behav Neurosci*, 13(2), 330-345.
- Carter, C. S., Braver, T. S., Barch, D. M., Botvinick, M. M., Noll, D., & Cohen, J. D. (1998). Anterior cingulate cortex, error detection, and the online monitoring of performance. *Science*, 280(5364), 747-749.
- Chaminade, T., Rosset, D., Da Fonseca, D., Nazarian, B., Lutchter, E., Cheng, G., & Deruelle, C. (2012). How do we think machines think? An fMRI study of alleged competition with an artificial intelligence. *Frontiers in human neuroscience*, 6.
- Dixon, S. R., Wickens, C. D., & McCarley, J. S. (2007). On the independence of compliance and reliance: Are automation false alarms worse than misses? *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 49(4), 564-572.
- Dzindolet, M. T., Beck, H. P., Pierce, L. G., & Dawe, L. A. (2001). *A framework of automation use*.
- Federal Aviation Administration. (2013). *Operational Use of Flight Path Management Systems: Final Report of the performance-based operations Aviation/Commercial Aviation Safety Team Flight Deck Automation Working Group*.
- Hesselmann, G., Sadaghiani, S., Friston, K. J., & Kleinschmidt, A. (2010). Predictive Coding or Evidence Accumulation? False Inference and Neuronal Fluctuations. *PloS one*, 5(3), e9926.
- Kiehl, K. A., Liddle, P. F., & Hopfinger, J. B. (2000). Error processing and the rostral anterior cingulate: An event-related fMRI study. *Psychophysiology*, 37(2), 216-223.
- Krach, S., Hegel, F., Wrede, B., Sagerer, G., Binkofski, F., & Kircher, T. (2008). Can Machines Think? Interaction and Perspective Taking with Robots Investigated via fMRI. *PloS one*, 3(7), e2597.

- Lee, J., & Moray, N. (1992). Trust, control strategies and allocation of function in human-machine systems. *Ergonomics*, 35(10), 1243-1270.
- Lee, J. D., & See, K. A. (2004). Trust in Automation: Designing for Appropriate Reliance. *Human Factors*, 46(1), 50-80.
- Madhavan, P., & Gonzalez, C. (2006). *Effects of sensitivity, criterion shifts, and subjective confidence on the development of automaticity in airline luggage screening*. Paper presented at the Proceedings of the Human Factors and Ergonomics Society Annual Meeting.
- Madhavan, P., & Wiegmann, D. A. (2007). Effects of information source, pedigree, and reliability on operator interaction with decision support systems. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 49(5), 773-785.
- Mathalon, D. H., Whitfield, S. L., & Ford, J. M. (2003). Anatomy of an error: ERP and fMRI. *Biological psychology*, 64(1), 119-141.
- McBride, S. E., Rogers, W. A., & Fisk, A. D. (2014). Understanding human management of automation errors. *Theoretical issues in ergonomics science*, 15(6), 545-577.
- Menon, V. (2011). Large-scale brain networks and psychopathology: a unifying triple network model. *Trends Cogn Sci*, 15(10), 483-506.
- Merritt, S. M., Heimbaugh, H., LaChapell, J., & Lee, D. (2013). I Trust It, But I Don't Know Why : Effects of Implicit Attitudes Toward Automation on Trust in an Automated System. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 55(3), 520-534.
- Meyer, J. (2004). Conceptual issues in the study of dynamic hazard warnings. *Hum Factors*, 46(2), 196-204.
- Mosier, K. L., Skitka, L. J., Heers, S., & Burdick, M. (1998). Automation Bias: Decision Making and Performance in High-Tech Cockpits. *The International Journal of Aviation Psychology*, 8(1), 47-63.
- Oliver, R. L. (1980). A Cognitive Model of the Antecedents and Consequences of Satisfaction Decisions. *Journal of Marketing Research*, 17(4), 460-469.
- Onnasch, L., Ruff, S., & Manzey, D. (2014). Operators' adaptation to imperfect automation – Impact of miss-prone alarm systems on attention allocation and performance. *International Journal of Human-Computer Studies*, 72(10–11), 772-782.
- Parasuraman, R., & Manzey, D. H. (2010). Complacency and bias in human use of automation: an attentional integration. *Hum Factors*, 52(3), 381-410.
- Parasuraman, R., Molloy, R., & Singh, I. L. (1993). Performance consequences of automation-induced 'complacency'. *The International Journal of Aviation Psychology*, 3(1), 1-23.
- Parasuraman, R., & Riley, V. (1997). Human and Automation: Use, Misuse, Disuse, Abuse. *Human Factors*, 39(2), 230-253.
- Ress, D., & Heeger, D. J. (2003). Neuronal correlates of perception in early visual cortex. *Nat Neurosci*, 6(4), 414-420.
- Rice, S., & McCarley, J. S. (2011). Effects of Response Bias and Judgment Framing on Operator Use of an Automated Aid in a Target Detection Task. *Journal of Experimental Psychology: Applied*, 17(4), 320-331.

- Sanchez, J., Rogers, W. A., Fisk, A. D., & Rovira, E. (2014). Understanding reliance on automation: effects of error type, error distribution, age and experience. *Theor Issues Ergon*, 15(2), 134-160.
- Shenhav, A., Botvinick, Matthew M., & Cohen, Jonathan D. (2013). The Expected Value of Control: An Integrative Theory of Anterior Cingulate Cortex Function. *Neuron*, 79(2), 217-240.
- Skitka, L. J., Mosier, K., & Burdick, M. D. (2000). Accountability and automation bias. *International Journal of Human-Computer Studies*, 52(4), 701-717.
- Snizek, J. A., Schrah, G. E., & Dalal, R. S. (2004). Improving judgement with prepaid expert advice. *Journal of Behavioral Decision Making*, 17(3), 173-190.
- Staudinger, M. R., & Buchel, C. (2013). How initial confirmatory experience potentiates the detrimental influence of bad advice. *Neuroimage*, 76, 125-133.
- Suen, V. Y. M., Brown, M. R. G., Morck, R. K., & Silverstone, P. H. (2014). Regional Brain Changes Occurring during Disobedience to “Experts” in Financial Decision-Making. *PloS one*, 9(1), e87321.
- Tanner Jr, W. P., & Swets, J. A. (1954). A decision-making theory of visual detection. *Psychological Review*, 61(6), 401-409.
- Wickens, C. D., & Dixon, S. R. (2007). The benefits of imperfect diagnostic automation: A synthesis of the literature. *Theoretical issues in ergonomics science*, 8(3), 201-212.
- Wickens, C. D., Rice, S., Keller, D., Hutchins, S., Hughes, J., & Clayton, K. (2009). False alerts in air traffic control conflict alerting system: Is there a “cry wolf” effect? *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 51(4), 446-462.
- Yaniv, I., & Kleinberger, E. (2000). Advice Taking in Decision Making: Egocentric Discounting and Reputation Formation. *Organizational Behavior and Human Decision Processes*, 83(2), 260-281.

## **BIOGRAPHY**

Kimberly S. Goodyear graduated from South Pasadena High School, South Pasadena, California, in 2001. She received her Bachelor of Science from San Diego State University, San Diego, California, in 2005.

1.

**1. Report Type**

Final Report

**Primary Contact E-mail**

Contact email if there is a problem with the report.

fkrueger@gmu.edu

**Primary Contact Phone Number**

Contact phone number if there is a problem with the report

703.993.4358

**Organization / Institution name**

George Mason University

**Grant/Contract Title**

The full title of the funded effort.

Neural Signatures of Trust during Human-Automation Interactions

**Grant/Contract Number**

AFOSR assigned control number. It must begin with "FA9550" or "F49620" or "FA2386".

FA9550-13-1-0017

**Principal Investigator Name**

The full name of the principal investigator on the grant or contract.

Frank Krueger

**Program Manager**

The AFOSR Program Manager currently assigned to the award

Dr. Benjamin A. Knott

**Reporting Period Start Date**

01/01/2013

**Reporting Period End Date**

12/31/2015

**Abstract**

Well-calibrated human-automation trust (HAT) is an essential ingredient for efficiency, communication, and safety in complex human-automation interactions. A dichotomy between HAT and human-human trust (HHT) has been proposed: some scholars argue that HAT and HHT are fundamentally different due to initial perception and lack of intention on the part of automation, while others claim that HAT and HHT are equal, since similar social interactions as between humans can be elicited when automation is designed to be human-like. Although, recent behavioral research has provided evidence for both accounts and a plethora of neural evidence for HHT already exists; however, the underlying neural signatures for HAT and its relationship to HHT are still unexplored. Behavioral measures alone are unlikely to allow one to distinguish between HHT and HAT, because the same behavioral outcome can be associated with very different underlying neural mechanisms. Assessing both performance and brain function can provide more information than either alone. The objective of this proposal was to investigate the similarities and differences of the neural systems of HAT and HHT in a series of three studies that combined a behavioral X-ray luggage-screening task with functional magnetic resonance imaging (fMRI) and manipulated reliabilities of advice (unknown to the participants) as the key feature for HAT and HHT interactions. Healthy participants were asked to search for knives hidden in densely cluttered X-ray images of luggage after receiving advice (presence or absence of a knife) from a human or automated luggage inspector

(framed as experts). HAT and HHT were measured as the acceptance rates of advice either giving by the machine or human agent. By adopting a comprehensive, interdisciplinary research program including scientists from social cognitive neuroscience, psychology, and human factors, we accomplished the overall objective of this proposal by pursuing the following three specific aims:

**Aim #1:** Neural signatures of HAT based on reliable human-automation interactions. In study 1, participants performed the security screening task and decided whether to search or clear the luggage after receiving advice from a human or automated luggage inspector with a manipulated reliability of 90%. HHT was initially lower than HAT, probably due to the preconceived notions of automation being perfect. However, over time differences between HHT and HAT disappeared based on a higher degree of confidence toward the human adviser to perform the task based on the received feedback. This reinforcement learning process was mirrored by activations in reward-sensitive brain regions, including the dorsal striatum and ventromedial prefrontal cortex. In summary, comparing HHT and HAT study 1 provided the first neural evidence showing how automation bias mediates these types of trust, thus leading to behavioral differences in the context of advice taking.

**Aim #2:** Neural signatures of HAT based on unreliable human-automation interactions due to high false alarm rates. In study 2, participants completed the X-ray luggage-screening task by either rejecting or accepting bad or good advice from either a machine or human inspector with a manipulated reliability of 60% (false alarm rate). Unreliable advice decreased performance overall. HHT was lower than HAT during bad advice, presumably due to reevaluation of expectations arising from association of dispositional credibility for each agent. Trust differences engaged brain regions associated with the mentalizing network for evaluating personal characteristics and traits (precuneus, posterior cingulate cortex, temporoparietal junction) and the salience network for interoception (posterior insula). Posterior insula and left precuneus were the drivers of the HHT network that were reciprocally connected to each other and also projected to all other regions. In summary, study 2 revealed insights into the neural underpinnings of HAT and HHT associated with unreliable advice utilization due to high false alarm rates.

**Aim #3:** Neural Signatures of HAT based on unreliable human-automation interactions due to high miss rates (60%). In study 3, participants performed the X-ray luggage-screening task by either accepting or rejecting good or bad advice from either a human or a machine inspector with a manipulated reliability 60% (miss rate) of. HAT decreased more than HHT over time, possibly due to high expectations of reliable advice from a machine and changes in attention allocation due to miss errors. Brain areas involved with the salience and mentalizing networks, as well as sensory processing involved with attention were less active for HAT as for HHT. The HAT network consisted of attentional modulation of sensory information with the lingual gyrus as the driver during the decision phase and the fusiform gyrus as the driver during the feedback phase of the task. In summary, study 3 expanded on the existing literature by showing how misses degrade HAT in comparison to HHT, which is represented in brain regions involved in salience detection and self-processing with perceptual integration.

The performed studies are innovative, because they were among the first directly to examine and compare the neural signatures of HAT (and its relationship to HHT) in the context of human-automation performance applying a multi-disciplinary approach. The findings have significant implications for society because of progressions in technology and increased interactions with machines. Moreover, those findings are relevant to the Air Force Office of Scientific Research's mission aimed at fostering innovative research and enhancing the Air Force's impact on policies and operations related to national security by investing in the discovery of the foundational concepts of trust building and trust calibration during complex human-machine interactions. Overall, the successful completion of this project resulted in two substantive project outcomes: first, a significant increase in our knowledge about the underlying neural circuits of HAT calibration during complex human-automation interactions and second, the laboratory results provide a methodology and rationale for exploring HAT in field research and for developing transformative novel theories and models.

## Distribution Statement

This is block 12 on the SF298 form.

Distribution A - Approved for Public Release

## Explanation for Distribution Statement

If this is not approved for public release, please provide a short explanation. E.g., contains proprietary information.

## SF298 Form

Please attach your [SF298](#) form. A blank SF298 can be found [here](#). Please do not password protect or secure the PDF. The maximum file size for an SF298 is 50MB.

[AFD-070820-035\\_PI\\_Krueger.pdf](#)

**Upload the Report Document. File must be a PDF. Please do not password protect or secure the PDF. The maximum file size for the Report Document is 50MB.**

[Final\\_Performance\\_Report\\_PI\\_Krueger\\_Plus\\_Attachment.pdf](#)

**Upload a Report Document, if any. The maximum file size for the Report Document is 50MB.**

## Archival Publications (published) during reporting period:

Findings of study 1 were submitted as an abstract to the 21st Annual Meeting of the Cognitive Neuroscience Society (Boston, MA; April 5-8, 2014):

Title: How automation bias influences human-human and human-automation trust: An fMRI study

Authors: Goodyear K, Bowman A, Chernyak S, De Visser E, Parasuraman R, Krueger F.

Findings of study 2 were submitted as an abstract to the Society for Social Neuroscience Annual Meeting (Chicago, IL; October 16, 2015):

Title: Comparisons of advice utilization during human and machine agent interactions: a functional magnetic resonance imaging and effective connectivity study

Authors: Goodyear K, Parasuraman R, Chernyak S, Madhavan P, Deshpande G, Krueger F.

The research effort for this project culminated in the production of one dissertation. In April 2006, Kimberly S. Goodyear will defend her dissertation entitled "The neural basis of advice utilization During human and machine agent interactions" to the graduate faculty of George Mason University in partial fulfillment of the requirements for the degree of Doctor of Philosophy Neuroscience. The dissertation includes the findings from study 1 and study 2 (see attachment). The PI of the research project will act as the Dissertation Director.

Moreover, a manuscript entitled "Advice utilization during human and machine interactions: an fMRI and effective connectivity study" based on the findings of study 2 is currently under review as an original research article in the journal "Frontiers in Human Neuroscience":

Authors: Kimberly Goodyear, Raja Parasuraman, Sergey Chernyak, Poornima Madhavan, Gopikrishna Deshpande, Frank Krueger

Author Contributions: K.G. and S.C. acquired the data for analysis. K.G., R.P. and F.K. contributed to the conception of the design. K.G., R.P., S.C., P.M., G.D. and F.K. contributed to interpretation of the data. K.G., R.P., S.C., P.M., G.D. and F.K. contributed to drafting of the work and revising it critically. K.G., R.P., S.C., P.M., G.D. and F.K. approved the final version to be published. K.G., R.P., S.C., P.M., G.D. and F.K. agreed to be accountable for all aspects of the work.

Abstract: With new technological advances, advice can come from different sources such as machines or humans, but how individuals respond to such advice and the neural correlates involved need to be better understood. We combined functional MRI and multivariate Granger causality analysis with an X-ray luggage-screening task to investigate the neural basis and corresponding effective connectivity involved with advice utilization from agents framed as experts. Participants were asked to accept or reject good or bad advice from a human or machine agent with manipulated reliability (high false alarm rate). We showed

that unreliable advice decreased performance overall and participants interacting with the human agent had a greater depreciation of advice utilization during bad advice. These differences in advice utilization can be due to reevaluation of expectations arising from association of dispositional credibility for each agent. We demonstrated that differences in advice utilization engaged brain regions associated with evaluation of personal characteristics and traits (precuneus, posterior cingulate cortex, temporoparietal junction) and interoception (posterior insula). We found that the right posterior insula and left precuneus were the drivers of the advice utilization network that were reciprocally connected to each other and also projected to all other regions. Our behavioral and neuroimaging results have significant implications for society because of progressions in technology and increased interactions with machines.

Finally, another manuscript entitled "An fMRI and effective connectivity study investigating miss errors during advice utilization from human and machine agents" based on the findings of study 3 is currently under review as an original research article in the journal "Social Neuroscience":

Authors: Kimberly Goodyear, Raja Parasuraman, Sergey Chernyak, Ewart de Visser, Poornima Madhavan, Gopikrishna Deshpande, Frank Krueger

Author Contributions: K.G. and S.C. acquired the data for analysis. K.G., R.P. and F.K. contributed to the conception of the design. K.G., R.P., S.C., P.M., G.D. and F.K. contributed to interpretation of the data. K.G., R.P., S.C., E.D.V., P.M., G.D. and F.K. contributed to drafting of the work and revising it critically. K.G., R.P., S.C., E.D.V., P.M., G.D. and F.K. approved the final version to be published. K.G., R.P., S.C., E.D.V., P.M., G.D. and F.K. agreed to be accountable for all aspects of the work.

**Abstract.** As society becomes more reliant on machines and automation, understanding how people utilize advice is a necessary endeavor. Our objective was to reveal the underlying neural mechanisms during advice utilization from expert human and machine agents with fMRI and multivariate Granger causality analysis. During an X-ray luggage-screening task, participants accepted or rejected good or bad advice from either the human or machine agent framed as experts with manipulated reliability (high miss rate). We showed that the machine-agent group decreased their advice utilization compared to the human-agent group and these differences in behaviors during advice utilization could be accounted for by high expectations of reliable advice and changes in attention allocation due to miss errors. Brain areas involved with the salience and mentalizing networks, as well as sensory processing involved with attention, were recruited during the task and the advice utilization network consisted of attentional modulation of sensory information with the lingual gyrus as the driver during the decision phase and the fusiform gyrus as the driver during the feedback phase. Our findings expand on the existing literature by showing that misses degrade advice utilization, which is represented in a neural network involving salience detection and self-processing with perceptual integration.

**Changes in research objectives (if any):**

None

**Change in AFOSR Program Manager, if any:**

Dr. Benjamin Knott replaced Dr. Joseph Lyons on August 1st, 2013 as the Program Officer for the Trust and Influence portfolio.

**Extensions granted or milestones slipped, if any:**

None

**AFOSR LRIR Number**

**LRIR Title**

**Reporting Period**

**Laboratory Task Manager**

**Program Officer**

**Research Objectives**

**Technical Summary**

**Funding Summary by Cost Category (by FY, \$K)**

	Starting FY	FY+1	FY+2
Salary			
Equipment/Facilities			
Supplies			
Total			

**Report Document**

**Report Document - Text Analysis**

**Report Document - Text Analysis**

**Appendix Documents**

**2. Thank You**

**E-mail user**

Mar 18, 2016 11:57:30 Success: Email Sent to: fkrueger@gmu.edu